DOCUMENT RESUME

ED 221 566 TM 820 589

TITLE Bilingual Evaluation Technical Assistance: Workshop Text II. Planning for Evaluation Design.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE [80]
NOTE 152p.; For related documents, see TM 820 588-592.

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS *Bilingual Education; Educational Planning; Elementary Secondary Education; *Evaluation Methods; Program Design; Program Evaluation; *Research Design; *Technical Assistance; Test Selection; *Workshops

ABSTRACT

This text is designed to accompany workshop instruction in developing an evaluation design for evaluating bilingual programs. It is addressed to those involved with bilingual programs generally, educators directly involved in planning and implementation, and research and evaluation specialists. It is organized into two separate workshop sessions: A: Basic Concepts in Design, and B: Selecting Evaluation Designs. Workshop A: Basic Concepts in Design, is an introduction to the concepts of threats to validity, controlled assignment and comparison, and the elements of design. It is designed for audiences with little experience in evaluation. Workshop B: Selecting Evaluation Designs assumes audiences are comfortable with the concepts and terminology in Workshop I and are now ready to consider the more technical and specific features of representative designs. This second workshop considers four planning questions: (1) What do I need to consider in selecting an evaluation design? (2) Which designs best fit the particular demands of bilingual education? (3) How can I adapt these designs to my particular setting? and (4) What other designs might be useful for special evaluation purposes? (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221566

Bilingual Evaluation Technical Assistance

Workshop Text II

Planning for Evaluation Design

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G Grey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 589

Bilingual Evaluation Technical Assistance

Workshop Text II

. Overview: Evaluation of Bilingual Programs

. Planning for Management

. Planning for Evaluation Design

How to Use This Text

Planning for Evaluation Design is designed to accompany workshop instruction in evaluating bilingual programs. Instructional activities to provide participants with practice on skills presented in this workshop can be found in Participant's Handbook III: Planning for Evaluation Design.

You will want to read this workshop text before participating in the actual sessions. You might want to complete the "Participant Preparation" form provided at the beginning of the Participant Handbook accompanying this text. Bring questions you have and examples from your own program to the workshop.

Planning for Evaluation Design addresses three levels of involvement with bilingual evaluation. At the most general level, we hope that all people involved in bilingual programs, both bilingual and monolingual program teachers, students, parents, and community members will find the material valuable. A second audience is educators directly involved in bilingual program planning and implementation who are familiar with many of the general principles of curriculum development and testing, but who have no special training in evaluation. The third group for

which the materials were designed is research and evaluation specialists who may not be well acquainted with the special demands of bilingual program evaluation.

Since individual sections of the text will be more relevant to one group than to another, the Table of Contents is keyed by section to the three audiences described above. Consult the key to identify those parts of the text that match your interests and level of expertise. The workshop text, Planning for Evaluation Design, is organized into two separate workshop sessions: A: Basic Concepts in Design and B: Selecting Evaluation Designs. Workshop A: Basic Concepts in Design is an introduction to the concepts of threats to validity, controlled assignment and comparison, and the elements of design. It is designed for audiences with little experience in evaluation.

Workshop B: Selecting Evaluation Designs assumes audiences are comfortable with the concepts and terminology in Workshop I and are now ready to consider the more technical and specific features of representative designs. This second workshop considers four planning questions:

1) What do I need to consider in selecting an evaluation design?
2) Which designs best fit the particular demands of bilingual education?
3) How can I adapt these designs to my particular setting?
4) What other designs might be useful for special evaluation purposes?

Table of Contents: Workshop A

Basic Concepts in Design

		Intended Audience	Page
Part I	The Role of Design in Bilingual Evaluation	G	1
	What is an evaluation design?		1
	Why use evaluation designs?		3
	When should you use designs?		6
	What kinds of questions can designs answer?		7
Part II	Preplanning Considerations	G	10
	Basic features of bilingual evaluation		11
	What can go wrong during the the evaluation?		14
	Avoiding threats to validity: the sample		15
	Selection		15
	Maturation		17
	Threats to validity: instrumentation		19
	Instrument validity		20
	Instrument reliability		23
	Administration procedures		25
	Threats to validity: extraneous factors		26

	Intended Audience	Page
Basic features of design		28
Comparison		28
Controlled assignment		29
Multiple design		31
Part III Planning Activities	GS	33
Identifying the necessary questions		34
Selecting the unit of observation		39
Selecting the measures		41
Setting a timeline		42
Analyzing the data		44
Workshop B: Selecting Evaluation Designs		48
Introduction		49
Factors influencing design selection		51
Similarities to other educational programs		51
Differences from other educational programs		52
Diversity		52
Assessment		52
Accountability		53
Conditions and events influencing bilingual programs		54
Selection criteria		56
Evaluation setting		56
Purpose of the evaluation		58

	Intended Audience	Page
Comparative data required		58
Design requirements		59
Relative advantages of designs		59
Recommended Designs		63
Introduction		63
Notation		65
Recommended designs for bilingual programs		67
Historical		59
Treatment specification		74
Criterion referenced		81
Norm referenced		85
Troubleshooting Design Problems		96
Special Purpose Designs		102
Intact group designs		106
Regression designs		110
Equivalent group designs		118
References		130

Audience Key:

G: Appropriate for all levels of expertise including parents and community

S: Teachers, administrators and school site staff with instructional and administrative experience in bilingual programs, but no specialized evaluation training

R: Evaluators and research staff familiar with design, measurement, and implementation procedures used in program evaluation, but unfamiliar with the special demands of bilingual program education.

Planning for Evaluation Design

Workshop A: Basic Concepts in Design

Bilingual Evaluation Technical Assistance
Workshop Text III: Planning for Evaluation Design

The Role of Design in Bilingual Evaluations

What is an Evaluation Design?*

What is an evaluation design? In the narrowest sense of the term, a design is a strategy which dictates from whom evaluation information will be collected, with what measuring device, and when data collection will occur. Specifically, when an evaluator plans a design, he or she determines the following:

- The groups or units from whom information will be collected
- The measurement instruments to be used
- The times when the instruments will be administered
- The available procedures for analyzing the data

Evaluation design is also a management tool for organizing those data collection activities. It is a strategy whereby one identifies the kinds of questions to be answered by evaluation, specifies which information best addresses these questions, and incorporates techniques to make this information as credible as possible given constraints imposed by setting or time. "Design" does not merely refer to the statistics used for data analysis, although data analysis procedures do affect the credibility of evaluation results. Rather it is the term for all those organizing activities related to information gathering: asking the right questions, identifying the necessary sources of information, selection of appropriate instruments, applying appropriate data analysis procedures,

*The conceptualization of evaluation design presented here is derived from Cronbach and Shapiro, 1978. The reader is referred to Designing Educational Evaluations, by Lee J. Cronbach with the assistance of Karen Shapiro, Stanford University, April 1978: Occasional Papers of the Stanford Evaluation Consortium.

and making warranted inferences. To design an evaluation you should have a firm understanding of the political context in which you will be operating, theories and assumptions held by program participants, and the program itself. You need to know what kinds of data gathering activities, from case studies through achievement tests, will yield information most relevant to the questions guiding your evaluation. You should be familiar with alternative data analysis procedures and the kinds of inferences they support. In short, design consists of a potpourri of skills and knowledge you need to make your evaluation credible.

This workshop focuses on the methodological aspects of design--identifying what is to be evaluated, asking the necessary questions, selecting your sample, and planning data collection at appropriate times. Some of these activities are dealt with in more detail in other BETA workshops.

Planning for Student Assessment presents guidelines for instrument selection; Planning for Implementation offers strategies for program description and data gathering related to degree of implementation; and Planning for Management provides techniques for scheduling and managing data collection activities. You will want to refer to these texts as needed.

Why Use Evaluation Designs?

One obvious reason for using an evaluation design is to impose efficiency on the way the evaluation is conducted. A well-thought-out design will keep your plans for collecting evaluation information from lurching along in a haphazard fashion. By basing your data collection strategy on a good design, you can be relatively assured that measurements will be taken at the right time, from the appropriate people, and with the right kind of instrument.

Evaluation designs, though, are much more than a management tool. In fact, the primary reason for using a good design is to enhance the quality of the evaluation information. When you employ a defensible design, you elevate your findings from the realm of "educated guesses" to the status of trustworthy information.

Repeatedly mentioned in this chapter is the notion of a "good" design. A "good" design, essentially, is one that eliminates competing explanations for observed results. In other words, a "good" design eliminates the possibility that results (such as declines or gains in students' test scores) are due to factors other than the program itself. The notion of eliminating competing explanations perhaps can best be conveyed through a concrete illustration of small-scale summative evaluation:

"Why Evaluation Design: A Story"

Mark M. submitted his final evaluation report on the bilingual program at Stevens School to the district's Office of Program Development. It had been a controversial program, and district officials wanted to know whether the project was worth the investment of time and money. One official even mentioned to Mark that he supposed students would learn "just as much without the program".

Mark knew that the evaluation results would be important in determining the future of the bilingual program at Stevens School. Consequently, he was encouraged when students' posttest scores showed considerable gain over their pretest scores. In his evaluation report, Mark concluded that the program was very successful and should be continued. He based his conclusions regarding the positive effect of the bilingual program mostly on the test scores and the analyses he had done.

Two weeks later Mark received a formal response to his evaluation report from the Office of Program Development. He was shocked. Officials in the office disagreed with his conclusions and criticized his methodology. In effect, they said that flaws in the way he collected the data and the way he carried out the analyses made the results invalid. Some of their specific comments were:

"The pretest-posttest interval was too short, and the students' gains were due to their familiarity with the test itself."

"Many of the students were involved in a special after-school English tutoring program. I bet the after-school program caused the gains, not the District's bilingual education program."

"The pretest scores were so low that it is natural to expect scores to go up during the course of the school year - whether or not the students took part in a special bilingual program."

"Why couldn't some other bilingual program do even better? In fact, how do we know this program did any better than no program at all?"

Needless to say, Mark was disappointed. His report was filed on a back shelf, and the final decision on the future of the bilingual program at Stevens School was made with almost no attention to his efforts and recommendations.

The End

The chief flaw in Mark's evaluation was his inability to eliminate competing explanations for the results. As described, the district officials found alternative reasons for the students' test score gains, reasons that had nothing to do with the bilingual program itself and reasons that could not be ruled out by the data that Mark had collected.

Situations such as the one experienced by the unfortunate Mark are repeated far too often. Better knowledge of evaluation design could have prevented this frustrating experience and could possibly have influenced decision makers to support the program in question. What is doubly frustrating is that the fundamental concepts and procedures of good evaluation design are not that complex, and the payoffs from employing these concepts can be great.

When Should You Use Designs?

Evaluation designs are techniques of systematic data collection. These techniques improve the plausibility of information, rule out completing explanations, and enhance the generalizability of findings over students, teachers, and programs. When should you use systematic data gathering procedures? Whenever you are concerned with providing information for decision makers and the decision makers need reliable, believable information.

The need for credible information arises in several kinds of bilingual evaluations. Decision makers may be interested in program context, inputs, implementation, or outcomes. The courts are often concerned with whether the program is promoting effective participation in the educational process (Lau V. Nichols, 1973). State and federal funding agencies need reliable data related to funds allocation (program inputs), program installation (implementation), and student achievement (outcomes). School district administration is concerned with student achievement and the effective allocation of educational resources. Project staff, in addition to having the above concerns, have a special need for information related to program improvement. Parents, whose participation is required in the planning and implementation of bilingual programs, are concerned both with student outcomes and the school environment. Thus, there are at least four kinds of evaluation efforts requiring the use of designs: context, input, implementation and outcomes.

What Kinds of Questions Can Designs Answer?

Each of the four evaluation foci on the previous page, context, input, implementation and outcomes, generates different sets of questions to guide the evaluation. If decision makers wish information about program context, they may have some of the following questions in mind:

What are the political reasons for program installation?

Who is interested in the program, and why?

What are community and staff attitudes toward the program?

What are the project staff's theories of bilingual education; do staff members concur in their theories?

These questions are designed to assess the "climate" in which the program is operating, possible conflicting interests in program results, and relevant conditions existing prior to program implementation.

Funding agencies, administrators, and project staff often wish to know about the quality and quantity of program inputs in order to assess the level of program implementation or the relationship of these inputs to program outcomes. Some input questions are:

Which languages are spoken by students and how well?

What is the home language usage of students?

How many bilingual/ESL teachers are available and what is their competence to teach in bilingual programs?

What kinds of materials are being used?

How much money is being spent on salaries, materials, aides, etc.?

These questions deal with the kinds and amounts of resources being used in the bilingual program.

Evaluation of program implementation is of especial concern to the evaluator. You must know that the program really exists, what its

goals are, and what it looks like, before you can plan data collection. Program implementation data is also very useful for substantiating theories of bilingual instruction, assessing future program needs, and examining the relationship between program participation and student achievement. Implementation evaluation focuses on several of the following kinds of questions:

Does the program exist?

What does it look like?

How much time is being spent in various activities?

What are teachers' preferred teaching styles?

How are materials being used?

What are the patterns of teacher-student language use?

Is the program complying with legal guidelines?

How can the program be improved?

A final evaluation focus is on program outcomes. This perspective is perhaps the most familiar and certainly the one that first comes to mind when you think of "evaluation." The focal point of outcomes evaluation is on student achievement. In fact, federal and state regulations require student achievement data. Evaluators, then, may be requested to provide information about:

How well does the program promote student achievement vis-à-vis a national norm group?

How do students in the program compare to those in a similar bilingual program?

How do students compare to students receiving no special bilingual instruction?

What is the pattern of student achievement over time?

To summarize, bilingual evaluators are interested in designs that provide credible information for a wide range of questions. The reason that so many different types of information are gathered in bilingual evaluations is that programs are often very new and there are many stakeholders in the evaluation. There is little baseline information avaliable for decision-makers to know what the program should look like when it is fully implemented, what desirable and normal patterns of language growth in the native language and English should be, and what instructional strategies and materials are most positively associated with desired student outcomes. Additionally, different audiences have different information needs. Thus, the bilingual evaluator has the task of providing a variety of background, descriptive, process, and outcome information to inform subsequent program planning and evaluation efforts.

Throughout this section the word "credible" has been used when discussing information collected during an evaluation. The next section of this chapter will explain threats to the credibility of bilingual evaluation findings and propose strategies of systematic data collection to counter these threats.

Preplanning Considerations

The purpose of this section is to take you through the preplanning considerations associated with the basic features of bilingual evaluation, the threats to the validity of the evaluation, and the basic features of design.

Basic Features of Bilingual Evaluation

In planning designs for bilingual programs, it is important to remember that while they share many features with monolingual programs, they also possess unique characteristics. Thus, the evaluator must simultaneously consider these similarities and differences while collecting data on such variables as student achievement, teacher preparation, and instructional strategies.

Foremost among the differences between bilingual and "regular" evaluation is the factor "language proficiency." One of the central concerns of the bilingual evaluator is language acquisition. How proficient are students in the first language? How is language taught? What external reinforcement is there for learning the second language and/or maintaining the native language? How does one measure language growth in a second language? What are reasonable expectations for language growth. Are these the same for different linguistic groups? What is the relationship between proficiency in the first and second languages? How do you define language proficiency for students having no written language? Instrument selection and interpretation of program results emerge from the evaluator's knowledge of theories of language acquisition.

Teachers participating in bilingual programs have different demands from those in monolingual settings. The level of language proficiency of the teacher, the teacher's attitudes toward the native language of the students, and her/his theory of bilingual education are variables that may affect program implementation and student outcomes. Documentation of these teacher characteristics may prove invaluable for interpreting evaluation findings.

A third major difference in bilingual programs is the kinds of instructional strategies used. The amount of time spent in the native and second languages, how this time is allocated and to which subjects, and the kinds and use of materials are often different from that in monolingual programs. Additionally, not as much is known about the effectiveness of various bilingual instructional practices. These differences may require that the evaluator gather information related to theory-building within the context of the evaluation. If you are asked to conduct small research studies, you will need to select designs that will allow you to make comparisons between "treatment" and "control" groups, in addition to the inferences required by the evaluation.

A final difference between bilingual and monolingual evaluation relates to the "state of the art" in bilingual education. Because many programs are fairly recent, there is little information available for judging program effectiveness. Current instruments are almost universally inadequate for assessing language growth in native languages; their relevance for assessing growth of second language learners is yet to be determined. The term "bilingual program" has many different meanings ranging from total immersion in English to the use of two languages to conduct instruction in all subjects. Program definitions differ both between and within projects. The political context of the program is an important variable absent in monolingual settings. Some communities welcome bilingual education, some are hostile, some indifferent or uninformed. These differences make bilingual evaluation a more complex task than otherwise.

You may have to develop your own instruments and interpret the results without clear guidelines. You may have to collect data related to factors such as community attitude, instructional techniques or teacher language proficiency simply to understand the most basic program outcomes: student achievement test scores. You should be aware that the kinds of evaluation questions addressed and their attendant evaluation designs will often be more extensive than those associated with the evaluation of monolingual programs.

What Can Go Wrong During the Evaluation?

In the story of Mark's evaluation, officials were able to find competing explanations for the gain in students' test scores. Basically, flaws in the data collection procedures threatened the validity of conclusions presented in the evaluation report. Apparently, Mark did not realize that students who repeatedly take the same test over a short period of time do become familiar with the test items. Their scores may have increased, not because they had experienced the program, but simply because they learned the test! This phenomenon is called "testing effect". In a good evaluation, the design for data collection will minimize or eliminate testing effects by properly spacing administrations of the test for pretest and posttest administrations. Had Mark applied these simple concepts, his findings may have been viewed in better light.

In addition to testing effects, there are several additional "threats" that can reduce the validity of an evaluator's conclusions. This section describes three sources of threats: sample selection, instrumentation effects, and extraneous factors. In order for you to be able to assert with confidence that program results are linked in some direct way to program experiences, you must be able to state three things that the unlucky Mark M. was not able to say:

- Program participants' background characteristics or development were not alone responsible for program results.
- The measures you used and how you administered them substantiate your inferences,
- Participants did not acquire the desired outcomes due to causes outside of the program.

This section looks at each area where you can run into trouble during your evaluation and suggests ways to avoid threats to the validity of your evaluation conclusions.

Avoiding Threats to Validity: The Sample

Two characteristics of the sample may hide program effects: their background characteristics and their developmental characteristics. If you include in the bilingual program treatment groups sample from a population that is defined to be different from students not offered the treatment, your results are threatened by selection bias. If, because of the age of your sample or the skill being tested, program results can be expected simply because of student development, maturation is a threat to validity.

Selection. Selection refers to the manner in which students are assigned to groups for evaluation purposes. Ideally, students participating in the program and students to whom they are being compared are drawn from the same candidate pool, then randomly assigned to the bilingual program or to no program. Thus, students both in and out of the program would have similar language background, socioeconomic status, sex and achievement characteristics. Potential systematic group differences would probably be evened out by the random selection process because each student had an equal chance of belonging to the experimental or control sample. Because the evaluator did not manipulate the selection in any way by assigning students with particular characteristics to a certain condition, the selection of the sample would not be "biased". You could infer that results obtained from this comparison were fairly generalizable to similar student groups.

Students, however, unlike coins, cards, and colored balls, usually cannot be selected for programs by random assignment. For this reason, the evaluator has to be especially attentive to the representativeness of the student groups upon which he/she is planning comparisons. That is, are comparison groups similar in all characteristics that may influence their performance on the measures used in the evaluation? These characteristics include age, sex, language background, socio-economic status, language proficiency, and prior achievement. If, for instance, one group is comprised of older students, or those students who are initially more skilled, then fair comparisons cannot be made.

Selection bias is a pervasive problem in the evaluation of bilingual programs funded through federal monies. Federal regulations forbid the withholding of programs from eligible students, thus eliminating the possibility for random assignment. Even when it is possible to form a program and a "non-program" group, other factors may diminish their comparability. For example, the groups may differ with respect to initial language proficiency, exposure to instruction in certain content areas, cultural values, and home environments.

When differences between comparison groups cannot be "even out" through random assignment procedures, other methods may be used. One such method is matching. Students are found outside the program who match program students as much as possible. For example, if you are evaluating a sixth grade Korean bilingual maintenance reading program, you will need to find an intact group somewhere fitting this description.

A second way to counter selection bias is to avoid making between group comparisons at all. Instead, you can get several measures of the

bilingual group's progress over time and plot their achievement trend. Although you may not be able to generalize about student growth in the absence of the program, these longitudinal studies are particularly useful because they avoid the inferences that are threatened by selection biases.

A third strategy for those who cannot find comparison groups is to compare students with a statistical norm group. Students in the bilingual program will be compared with students in a reference group of interest to the decision makers. These reference groups may include the national norming sample of a standardized test, local norms, or norms established for certain language groups.

Maturation. During the course of the school year, students do grow older. Some of the improvement in their ability to perform certain tasks is a result of this natural growth, or "maturation". For example, when first graders begin trying to print, their small motor coordination skills are not fully developed. Some of their motor development occurs naturally as a function of their growth and has no relation to the writing instruction in school. Within a period of a year, a young child might make tremendous gains on a test of printing ability, even without instruction in school.

Maturation is an ever-present threat in evaluations of bilingual programs, particularly in the area of language assessment. It is widely acknowledged, for instance, that cognitive development and language competence are highly related to each other. As a child's level of cognitive functioning changes as she/he grows older, so does his/her language development, regardless of formal instruction in language received at school.

The rate of children's maturation further complicates conclusions about program's effects. For a variety of reasons--previous instructions, home environment, health--children differ from each other in the rate of their growth in language competence. It has been empirically observed, for example, that students who score in the lower quartile of standardized, norm-referenced tests tend to learn at a slower rate than students scoring in the upper quartiles. When students are placed in a bilingual program on the basis of low standardized test scores, their slower rate of maturation may lead to conclusions, such as "The program was ineffective because the students are still behind their peers". In actuality, the students may be progressing, but simply at a slower rate. In fact, low scoring students might have fallen further behind their peers had there been no bilingual educational intervention at all. Bilingual programs showing maintenance of achievement scores, rather than growth, may in fact be producing net gains, given the maturational patterns of low scoring groups.

One way to minimize the effects of maturation is to base evaluative judgments on comparisons such as those described in the previous section. The evaluator can compare the students in the special program with a similar group of students, a "control group", matched group, or norm group who do not receive the special program. The assumption is that both groups of students will have similar rates of maturation which will be reflected in their test scores. A second assumption about "maturation" is that for both bilingual and comparison group students, growth in cognitive skills and language proficiency is linear; they both have a constant rate of gain over time, beginning from their initial level of

of functioning. If either of these assumptions is not true, the use of comparison groups cannot be used to control for selection bias. If, however, you have evidence that growth is linear for both groups, you can then compare students in the bilingual program with those in the comparison group. Students in the bilingual program then would have to perform better than the comparison group in order to demonstrate that their growth is attributable to something (hopefully the program) other than normal maturation. If the students in the special program out-perform the comparison group, an evaluator can be reasonably certain that maturation was not responsible for the positive gains.

Threats to Validity: Instrumentation

"Instrumentation" is a general term for data collection strategies used during your evaluation. Instruments may refer to archival records, observations, interviews, standardized tests, classroom products, or questionnaires. If the data collection procedures are of poor quality, that is, not valid or reliable, of if they are administered improperly, then you will have difficulty using these data to support assertions about your evaluation results.

Although instrumentation may take many forms, all data collection methods have certain properties in common, they sample behavior in some way. This sample is used to make inferences about the domain of behavior sampled. The Congruence between the domain of behavior selected to make inferences about the bilingual program and the objectives of the bilingual program is necessary for you to have an unbiased estimate of the kinds of behavior students in the program are expected to exhibit. The issue of test bias is discussed fully in BETA Workshop III: Planning of Student Assessment.

Instrument quality. The quality of your instruments has two dimensions: test substance and test stability. Test substance is concerned with what behavior is sampled and what kinds of inferences this sampling allows you to make. Test stability refers to the accuracy of the instrument, the certainty with which we can assert our inferences. How well a test actually measures what it proports to is its validity. The consistency of test responses across similar test items and/or over repeated administrations is the test's reliability.

Validity. Threats to the credibility of your evaluation stemming from poor instrument validity generally happen because you either choose the wrong behavior to sample or you make incorrect inferences. For example, you are in charge of an evaluation to determine if a particular reading program is an improvement over the "look-say" method. For your measure, you select a standardized reading test that samples a broad number of skills common to programs throughout the nation. Test results indicate that pupils did not improve significantly from pre to posttest; you inform the project director that the reading program is "no good". This inference is threatened by instrument invalidity. The test you selected was not, in and of itself, enough of a sample of program specific reading skills for you to conclude that the program was not working. (Although you might want to say that insofar as a means of promoting these broad skills, the program may not be particularly relevant.)

As you see in the example above, test validity is a function of test use. And, there are many ways you can inadvertently misuse a test. First, you must select a test that is of appropriate difficulty for the group being tested. Standardized achievement tests generally contain items with content and difficulty appropriate for a specified sample of

students. Read the test manual to determine just who these students are. If your pupils differ, however, from the national norming sample, you may find that in spite of your best attempt to match test levels with grade levels, many students can still answer only a few test items. These scores, in fact, may simply be lucky guesses. In this instance, matching test difficulty to group achievement is more difficult; you cannot simply select the designated level of test. Instead, you may have to try a few forms at lower age/grade levels until you find one appropriate for your students.

While administering a lower or higher level test than recommended will take care of the problem of matching students' achievement to test difficulty, this practice creates a second problem in test use: out-of-level testing. If you do use an out-of-level test, one that is either above or below the level recommended by the test publisher, your scores may be difficult to interpret. This is because the norm group upon which the test score interpretations are based will be very unlike your students in age and achievement patterns. Thus, the scores have no comparative meaning. Another problem with out-of-level testing is that the school curriculum and the content of the test may not match. For example, you have been told that the majority of 6th grade students in the bilingual program have a "3rd grade reading level". You then select a reading test form for 3rd graders. Not surprisingly, the content of the 3rd grade test does not match what the 6th graders had been learning in class. The test then yields little information about what the students have learned with regard to the 6th grade reading curriculum. Often test publishers recommend "functional level" testing, that is, out-of-level

testing, so that the distribution of your students' scores will not be unusual or skewed. While out-of-level testing will give you score distributions requisite for certain statistical tests, you certainly aren't getting a measure of what students are being taught. Out-of-level tests should not be used if it results in a mismatch between test content and instruction.

Another validity threat often encountered in bilingual program evaluations centers on the use of norm-referenced score interpretations. It is common practice to estimate how well bilingual program participants would have performed in the absence of the program by comparing their growth with that of a norm-group. Estimates of achievement based on the use of normed scores may lead to wrong conclusions, though, if norm group characteristics do not resemble the characteristics of the bilingual program participants (see "Selection" in the previous section). Although minority, low income, and non-English speakers are "adequately represented" in norming samples according to their proportion in the U.S. population, there are no special norms available for each of these groups separately. Thus, comparisons of interest, your bilingual students compared to other bilingual students, for example, are not provided by current norm-referenced tests.

Threats to validity from instrumentation are best dealt with by careful selection of tests, supplemented by other data collection procedures. Evaluators ought to be certain that there is a close match between content and skills covered in the test and those emphasized in the program. The instruments should have been sufficiently validated and possess an acceptable level of statistical reliability, (treated below). And, as mentioned, evaluators should carefully scrutinize the manner in which a

test was normed, paying particular attention to characteristics of the norm group. Please note that criteria for selecting student achievement tests are treated in more depth in Workshop III: Planning for Student Assessment. The validity of measures other than achievement tests is discussed in Workshop IV: Planning for Implementation Evaluation.

Instrument reliability. A second dimension of instrument quality is instrument reliability. Unreliable measures produce scores that change markedly, or fail to change at all, between test occasions, simply as a matter of measurement error. A reliable instrument is also an accurate one. If the amount of measurement error associated with a certain testing procedure is great, then the score as an estimate of achievement is inaccurate. This difference between the test score and some "real" estimate of student performance may be so large as to make decisions about that individual impractical.

Even if instruments themselves have high reliability statistics for the norm sample, the test statistics may change for your particular students. Reliability and validity coefficients, as well as item difficulties and discrimination indices, are sample dependent. They describe the characteristics of a test for the standardization sample. If you are testing students who differ in noticeable respects from the standardization sample, either because of instructional history, language proficiency or cultural differences, you will need to recalculate test statistics for your own sample in order to get accurate instruments of test reliability and item characteristics.

A final aspect of test score reliability has to do with the common practice of using one test score, usually at the high or low end of the

score distribution, to define the group of students to receive special educational treatments. If you are selecting students for the bilingual program simply because they received an extremely low score, say 15th percentile or below on an English achievement test, and these students are similar in ethnicity, sex, and other factors to the school population taking the test, then these extreme test scores are subject to regression effects. When students are selected from a relatively homogeneous student population to participate in a program on the basis of a single extreme score, there is a naturally occurring tendency for these scores to move toward the center of the distribution on a second test occasion. Students who score near the middle tend to have more stable scores that will not fluctuate as much on a second test administration. This statistical regression happens because chance and measurement error operate more on the "lucky" and "unlucky" extreme scores, than on those clustered near the middle. Regression effects are not a problem if the low scoring students come from a different population than students who tested too high to be included in the program. It is also not too serious a problem if multiple selection criteria are used for pupil program placement.

Regression effects can have several implications for bilingual program evaluation. Students receiving the lowest pretest scores will probably show the greatest "growth" in scores on the posttest, due to regression effects alone. Evaluators might erroneously conclude that the program helped the poorer student better than the higher achieving students. However, if a program actually did help the lower achieving student more than the higher ones, then an evaluator might reasonably

attribute the effect to regression, and not the program. Catch 22!

Unreliability poses a special threat in the field of bilingual program evaluation. The need for tests in many different languages, covering a range of skill areas, has resulted in a flood of new measures which may be marketed without sufficient technical information obtained in appropriate examinee groups. One approach for reducing the threat of unreliability is to make judgments on the basis of multiple, rather than single, measurements. Test scores, for example, can be augmented by other indices of achievement such as grade, informal observations, work samples, and teacher judgments. By including multiple sources of information, the evaluator minimizes the risk of drawing conclusions on the basis of a measure of inadequate technical quality.

Test administration procedures. The use of tests and other data collection techniques that are reliable and valid goes a long way toward making conclusions drawn from these instruments valid. However, the use of high quality instrumentation does not ipso facto guarantee that you have dealt with all possible validity threats associated with instrumentation. The timing of your data collection can affect the credibility of your results. If you schedule a pretest and posttest too close together in time, the students' test performance may reflect the influence of remembering the test items themselves rather than the acquisition of skills.

In bilingual program evaluations, testing effects are likely to occur when students are tested in their native language and in English on the same subject matter. When such a testing practice is followed, students actually are taking the same test twice, once in their native language and the second time in English.

As mentioned earlier, testing effects can be minimized by spacing test administration apart in time, preferably at least six months. Using different forms, which are parallel in skills covered but which differ in the specifics of item content, is another technique to minimize testing effects.

Another consideration related to the timing of tests is concerned with empirical norming dates. If you are using a standardized, norm-referenced test, you should administer it within two weeks of the actual date that the test was given for norming purposes. If, for example, the test was normed in October but you pretest in early December, test score comparisons between your sample and the norm group are likely to be highly suspect. Your students have had two months more instruction and maturity than the norm group. When you are using the same test for pre and posttest purposes, norms should have been gathered at two intervals, fall and spring. Should the test you select not have data from both periods, the norm tables are probably estimates or extrapolations which should be taken into account in using them for comparative purposes.

Threats to Validity: Extraneous Factors

During the course of a program, it is quite likely that non-program events will have an effect on the students' performance. For example, an after-school recreation program may have a positive effect on students' attitudes toward school which, in turn, may result in better attention in class. It would be unfair to attribute students' better attention in class to the academic program if, in fact, a large share of their enthusiasm for school was actually due to the recreation program.

Extraneous threats, that is events that are confounded with program implementation, are especially likely to occur in schools where more than one special program is operating. Schools with federally funded bilingual programs, for example, are likely to be eligible for funds for other compensatory education programs, such as those supported by Title I legislation. When students are simultaneously enrolled in several compensatory programs, evaluators may have difficulty in attributing gains in the effects of one program or another.

As in the case of maturation, the most direct way to control such confounding effects is to employ a comparison, or control, group. The control group should consist of students who are subject to all of the same conditions as the students in the bilingual program, except, of course, for the bilingual program itself. If the bilingual program group performs better than the control group, the evaluator could legitimately conclude that the program was at least partially responsible for the positive results.

How Can I Reduce Threats to Validity?

In the preceding discussion on threats to validity, you have seen that a number of factors exist which, if not taken into account in the planning of evaluation, can threaten conclusions of the evaluation effort. A skeptic can rightfully attack the results of a poorly designed study if it does not rule out for alternative explanations of results.

The next section of this chapter provides guidelines for planning designs that avert many of those threats.

Basic Features of Design

The preceding two sections told you of the many problems you might encounter in conducting a bilingual evaluation. Preplanning considerations related to the difference between bilingual and monolingual evaluation and the threats to validity posed by sample, instruments, and extraneous factors in bilingual settings might make it appear that it is impossible to conduct a credible bilingual evaluation. But, it is exactly to meet these problems that certain features of design were created. Three powerful concepts will enable you to counter many treats to validity and to engage in the kinds in the data collection activities most appropriate for bilingual education. These concepts are: comparison; controlled assignment; and multiple designs.

Comparison. One logical way to counter many of the threats to validity described in the preceding section is to base your conclusions on comparative data, specifically by making comparisons between the program group andsome external standard. The simple notion of comparison is remarkably powerful and almost singlehandedly deals with all of the threats to validity contained in sample, instrumentation, and extraneous factors.

Under ideal circumstances, this is what an evaluator might do. First, the evaluator would find a group of students who were alike in all relevant educational characteristics, such as abilities, parent and home characteristics. Then he/she would assign a sample to a special program, or "treatment", while other samples continue in the regular educational program.

For example, the group under study may participate in a new intensive reading program designed to improve reading comprehension. Other than the intensive reading program, everything about the two groups was the same.

An evaluator might ask the following question about the reading program: Do students in the reading program show greater growth in reading comprehension than students in the regular program? To answer this evaluation question, the evaluator might give both groups a reading comprehension test at the beginning of the year and, finally, administer the test again at the end of the year.

Suppose that the control students' scores average 23 on the pretest and 29 on the posttest, while the reading program students' scores went from 23 to 40 on the posttest. The differences in scores between the two groups would be powerful evidence for the value of the reading program.

Unfortunately, it is never possible to find identical groups of students who differ only with respect to a single educational treatment. Yet, if judgments about the relative performances of two groups are to be fair, the groups must be as similar as possible. What should the evaluator do to enhance the comparability of the groups? One answer is controlled assignment.

Controlled assignment. As mentioned earlier, it is usually impossible to find two groups who match on all of the relevant educational features. Realistically, if you were to look at two fifth-grade classrooms and compare students' reading levels, math skills, parent educational level and so on, you would discover that the two classes are not identical.

Since evaluators cannot really find groups with identical characteristics, they may use a process which insures that both groups have similar characteristics and resemble, with only chance variations, the

general diversity which commonly occurs in the population they represent. This procedure is called random assignment. The notion behind random assignment is that any member of the population (such as second graders in a particular school) has an equal chance of being selected to be in one of the groups (either the control group or the treatment group).

Practically speaking, it is not easy to achieve randomization in a school setting. Students ordinarily are not assigned to classrooms on a random basis. Even within classrooms, students typically are assigned to special treatments on the basis of need, not through randomization. Indeed, it is hard to imagine a school situation where it would be possible to assign students to a special program by drawing names from a hat or using a table of random numbers! The situation is further complicated in federally funded bilingual programs where all students eligible for the program must be served, effectively precluding the use of any control or comparison group.

The concept of controlled assignment, however, effectively addresses the problem of finding an appropriate comparison group for the evaluation. If the process by which groups are formed is known, for example, then actual differences between the groups can be <u>controlled</u> or explained. As mentioned above, randomization controls assignment of students to groups by actually manipulating the grouping process. Nevertheless, an evaluator can use naturally occurring equivalent groups for comparison purposes. An example of naturally occurring equivalent control groups is a comparison between students enrolled in a bilingual program in one school and students enrolled in a similar bilingual program in another school in the same district. Or, you might find an "equivalent" comparison

group by locating another classroom in another school or district whose students have the same language proficiency levels, native language background and socioeconomic status as those enrolled in the bilingual program. Randomization, matching, and selection of comparison groups on the basis of a cutting score are all instances of controlled assignment. Any controlled assignment procedure can be applied to different kinds of groups, to students, to classrooms (without regard to how the students in them got there), to schools, or to districts. If the process by which your comparison groups were formed is known, and the potential systematic differences between them documented, you will be able to make fairly generalizable inferences about student behavior based on these comparisons.

Multiple designs. The third powerful concept for improvement of design is the use of multiple designs. We have seen the information needs of bilingual programs are extensive and the kinds of evaluation required, several. No one design can hope to address all evaluation needs. Remember that design refers to your preselected units of observation (students, teachers, classes, administrators), measures, the timeline for data collection, and your data analysis plans. It is inconceivable that any one particular combination of timeline, sample, instruments, and analysis would be appropriate for answering several different evaluation questions. For example, your evaluation design for gathering information to address the question, "Is the program being implemented as planned?" will differ from one that answers, "How well are students who learn to read in their native language doing when compared to those instructed only in English?" In the first instance,

You will need observation and questionnaire data about teachers collected at frequent intervals throughout the year. Your data analysis will consist, in part, in matching what you have discovered happening in the program to the official course description as stated in the funding proposal and/or course of study. For the second question, which may be asked during the same evaluation, you will need achievement test data on English reading and reading in the native language as well as some of the descriptive information gathered above. These data will probably be collected at the beginning and end of the year from students enrolled in the bilingual program as well as LES/NES students not receiving reading instruction in their native languages. You will probably analyze the data by testing for significant differences between group norms on the posttest.

We have seen that several data collection plans can be incorporated into the evaluation, each chosen to address information best gathered from a particular group by a particular method at a specified time. Some of these designs may include random assignment within the program in order to answer research-type questions. Other designs will involve neither comparison nor controlled assignment because they are intended to gather information to be used only within the confines of the particular program.

Although no one evaluation can hope to address all aspects of bilingual programs, the use of multiple designs will allow you to investigate different aspects of the program and even different facets of the same evaluation question.

Planning Activities

One of the underlying principles of evaluation design presented in the previous section is to plan sampling, measures, and data collection times so as to minimize or eliminate plausible competing explanations for your findings. The aspects of design that you will employ, comparison, controlled assignment, or multiple data sources, will vary depending upon the focus of your evaluation and the questions you plan to answer. In planning your evaluation design, then, you should ask yourself five questions:

1. What do I want to know? What information do I need?
2. Who is best able to provide that information?
3. How can I obtain the information?
4. When will I collect the information?
5. How will I examine the information?

These questions imply a five step planning process described in this section: 1) Identifying the necessary questions, 2) Selecting the units of observation, 3) Selecting the measures, 4) Setting a timeline for data collection, and 5) Analyzing the data.

Identifying the Necessary Questions

Although the evaluator is not solely responsible for identifying the questions that the evaluation is designed to answer, in fortuitous circumstances, she/he will be able to participate in the planning of the evaluation. When you do have the opportunity to meet with the stakeholders in the evaluation, your role will be to help them clarify the issues to be addressed, point out the kinds of information related to those issues that evaluation can and cannot provide, and suggest areas of investigation that may be useful to the program, but were overlooked by the decision-makers.

As an evaluator, you should ask the following planning questions:*

What is the purpose of the evaluation?

How will the information be used, and by whom?

What will we know after the evaluation that we don't know now?

What will we do after the evaluation that we can't do now for lack of information?

The answers to these questions will provide you with the substance and the intent of the evaluation. You will need to know both in order to plan your design.

An example will illustrate the importance of your familiarity with

---

*These questions are suggested by Michael Patton (1978) and represent a view of evaluation concerned with providing information to decision-makers rather than reaching definitive decisions.

the questions guiding the evaluation.

Why Identify the Necessary Questions?: A Story

Mark M. was hired in May, two months before the final report for Project Bilingual's second year of operation was due. He read the proposal, which was funded through Title VII, in order to find out the objectives of the evaluation. From his knowledge of Title VII regulations, he decided to use the following design: Give a standardized reading and math test to all students in the program, and look up their pretest scores on the cumulative folders. He selected the same test that was given in the fall, but an alternate form. In addition, he had teachers complete a questionnaire asking what materials were used, how many minutes of reading and math instruction were conducted in the native language, and how many in English, and what in-service training they had received during the year. Since the Title VII regulations required information about student achievement as compared to "what the students would have achieved without the program," he was sure that using pre-post growth comparisons with the standardized test norm group would meet this criterion. He also knew that Title VII regulations required that the evaluation "be useful in helping with program implementation efforts." Information about the use of materials, patterns of language use, and teacher preparation would surely satisfy that requirement, he thought.

Mark personally was dedicated to a utilization-focused view of evaluation. He really wanted his results to be used by Project Bilingual participants. Furthermore, learning from his earlier experience,

he knew that his results showing pupil gain were unassailable.

Two weeks later Mark dropped by the district office to find out just how useful his report had been. He was shocked. Officials in the office, while agreeing with his conclusions and finding his methodology impeccable, were singularly unimpressed with his finding. Some of their specific comments were:

"You found out what materials were available, but we don't even know <u>if</u> they were used or <u>how</u>. How can I make inferences about the usefulness of the materials without that kind of information?"

"Sure, the teachers listed the courses they took for professional growth. But do you know what Cross-cultural Sensitivity 101 at Twerp College consists of? Well, it's a series of visits to restaurants and nightclubs located in the Puerto Rican, Japanese, Filipino, Chinese and Samoan sections of town."

"What parts of the program are working? How can we make them work better?"

"Is the program beneficial to student attitudes toward school?"

"How do the parents feel? We have quite a commitment to community support at this school."

Needless to say, Mark was disappointed. His report was filed on a back shelf, and the final decisions about how to improve the bilingual program at Stevens School was made with almost no attention to his efforts and recommendations.

The End

The chief flaw in Mark's evaluation, <u>this time</u>, was his lack of communication with the stakeholders in the evaluation, those people who were potential users of the evaluation results. What he could have done was to talk to the project director, district administration, parents and

teachers to find out not only what they wanted to know (which would probably include the same type of information Mark had identified), but how it would be used. If he had known, for example, that program components and their relationship to student achievement were important to the district in identifying program strengths and weaknesses, he could have used the teacher questionnaire data in conjunction with the test results to pinpoint some relationships between classroom practice and curricular outcomes. By finding out the district was concerned with parent attitudes, he could have made an attitude questionnaire for them or interviewed parents. It would have been no difficulty to give students an attitude survey with the posttest. And, in drawing his conclusions about teacher proficiency based on the inservice program, knowing that food, drink and dance constituted a large part of it, this could have changed his recommendations.

In identifying the necessary questions, the evaluator has a two stage task: 1) to identify and organize the relevant decision makers so that he can find the information needs of specific groups, and

2) to share with the decision makers' choices about the nature, purpose, content, and methods of the evaluation.

Once the evaluator has identified the evaluation stakeholders and is involved in the process of directing the focus of the evaluation (not always a possibility), she/he needs to know two kinds of information: what is to be the major substance of the evaluation; program context, program inputs, implementation, or outcomes? How will information

be used--for formative, program improvement, decisions, or for summative, program worth, decisions?

In bilingual evaluation, the substance of the evaluation not only will depend upon the funding requirements and wishes of the decision makers, but also on the stage of program implementation. In the early years of program development, context and implementation areas are important. You have to know that the program exists and what it looks like before you can look at its outcomes. Secondly, it takes from three to five or more years for programs to form, and even longer to plot language development.

Outcomes will be looked at in all evaluations for two reasons: funding agencies require student achievement information, often reported in terms of growth vis-à-vis a comparison group of some kind, and in order to identify areas for program improvement, student achievement is one good criterion of program success. Input evaluation is helpful for funders primarily concerned with resource allocation, for districts trying to make budgetary decisions, and for testing hypotheses about quantity and types of inputs and their relationship to desired program outcomes. Of course no evaluation can attend to all four areas. The implications of each, however, should be made clear to those who commission the evaluation.

The intent of the evaluation, like the substance, will depend not only on decision maker interest, but the stage of program implementation. For beginning programs, formative (program improvement) information is most useful. Early on, teachers and other staff need help in identifying those parts of the program that need improvement. They also need strategies for

making instruction better. Attitude information is especially important in the early stages in determining the direction of program growth.

After the program has "settled in," district officials may want to know if "we are getting our money's worth," or "if the program is any good." After you have determined that the program is being implemented as intended, you will be asked to focus on summative (program effectiveness) questions.

To summarize, before you can plan your evaluation design, you will need to identify the questions guiding the evaluation. Whether these questions are formative or summative will depend on the interests of the stakeholders in the evaluation and the stage of program development. The focus of these questions, context, inputs, implementation, or outcomes will depend upon these same factors. Figure 3-1 summarizes some questions typical for each kind of evaluation and for formative and summative purposes.

Selecting the units of observation. One of the first considerations in planning your evaluation design is which groups are to be involved: Students? Classrooms involved in the same kind of program? Districts? It is no trivial matter to decide exactly what level of analysis is of primary interest in the evaluation. If results are to have implications for programs, then most likely you will be interested in several classrooms housing the bilingual program. If the program is fairly uniform throughout schools in any one district, you may wish to sample from schools in the district, rather than classrooms in a school. You may even be asked to provide information about several units: students,

schools and districts. In any event, you will have to plan your sampling strategy so that you have included those units about which you will be asked to generalize.

One you have decided which units to sample, you will have to devise a plan for insuring that the groups you have chosen are representative of the kind of people or units over which you wish to generalize. Traditionally, representativeness of units is assured by random assignment: pupils or classrooms were randomly assigned to "treatments" (program or no program) or intact "equivalent" groups were located. For bilingual programs, however, random assignment of students to the program or the location of equivalent intact groups are difficult to accomplish.

(1) In federally funded Title VII, and most state bilingual programs, assignment to the program is mandatory for those students who qualify. Thus, random assignment procedures cannot be used.

(2) Matching of students enrolled in bilingual programs with students not enrolled in a bilingual program is most difficult. Within the bilingual program, there is often a diversity of native language groups as well as levels of proficiency. It is unlikely that you will find this same pattern of proficiency in an intact classroom outside the program; if you do, the school is probably not in compliance with state or federal legislation. Furthermore, language proficiency is often confounded with socioeconomic status. Thus, while two groups may superficially appear to have the same pattern of language proficiency skills, different SES levels make comparisons between them suspect.

(3) Finally, the number of intact groups that must be found for statistically valid inferences about program versus "control" groups is often prohibitive in terms of time or money.

The solution often arrived at to the near-impossibility of controlling assignment through randomization is to control assignment by gathering as much information as possible about any comparison. This information can be used to explain resulting differences between the groups and to eliminate these systematic differences in the statistical analysis.

Selection of measures. Once you have selected your sample and established what kinds of comparisons you wish to make, equivalent group, norm-group, or within group, then you must specify what methods you will use for collecting information from the sample. These methods are generally referred to as "instrumentation." Instrumentation includes formal paper-pencil tests as well as case studies, observation, questionnaires, interviews, and the like. The difficulty in selecting appropriate instruments for bilingual evaluation has been alluded to throughout this workshop. These problems are discussed more fully in BETA Workshop Text III: Planning for Student Assessment. In spite of the "problems" with current methods of measurement, you will be asked to select systematic observation procedures. Three key concepts should help you avoid some of the threats to validity posed by instrumentation:

(1) Select only those instruments that provide the appropriate comparative information related to your evaluation questions.

(2) Select instruments with the best technical quality (reliability and validity) available.

(3) Use multiple methods for collecting data.

Setting a timeline for data collection. The third component of your evaluation plan is your schedule for data collection. If you have been called in to evaluate the program after it has begun, your time schedule will differ from one that coincides with the beginning of the project. The principles guiding your data collection plan, however, are the same: You must collect some data early in the program that reflects what students bring with them to the program. You must collect some data near the end of the program that reflects potential program impact.

The timing of the pretest, when you have a choice, is determined by two factors: (1) how early you are included in on the planning of the evaluation and (2) the type of test you are using. If you are hired before the program has begun, you will pretest after the students have been selected for the program but before instruction has started. Often a pretest score is one of the criteria used to place students into a program. If this is the case, you will be able to use this score as a pretest for your evaluation. If you plan to analyze data using regression models, when students have been assigned to the program on the pretest score alone, and they are similar in other respects to students not in the program except for this pretest score, you will not be able to use this score. It is better to schedule a pretest for this selected group at the beginning of the program.

By the same reasoning, posttesting at the end of the program, or at the end of instruction, will allow students to get the maximum benefit from the program.

If, however, you are using norm-group comparisons for your evaluation design, you must pre-and posttest within two weeks of the empirical norming dates of the test you use. Norm group scores are obtained at specific times during the year. Scores taken at times different from empirical norming dates are not comparable because the norm tables for periods between these norming dates are estimated. Thus, strict adherence to the time band that includes two weeks before and two weeks after the midpoint of the period when the normative data were collected is a safer practice when making norm-group comparisons.

A final consideration in the timing of data collection is that of student mobility. For example, if the program is a two or three year project, the posttest will be administered at the end of the period, although interim measures will be taken. If during this period many students have exited from the program or have moved away, either these departed students must be found and posttested or conclusions about program effects must be confined to the kinds of students who have not moved. Posttests should be timed to "catch" as many students as possible for whom you have pretest scores.

Analyzing the Data

You should consider techniques for examining data as part of your evaluation plan. Although data analysis techniques do not determine what instruments you will use or how, both of these aspects of the design are influenced by the requirements of certain data analysis procedures. You would not plan your evaluations by saying, "I am going to look at the correlation between the number of students who do well on the reading posttest and the number of minutes of instruction in Farsi." Neither would you say, "I am going to do a two-way analysis of variance between treatment and no-treatment classrooms and students with scores above and below 50 on the pretest." Instead, you will think about what kinds of things you will be asked to say about the program. Will you have to make comparisons between program students and students not in a bilingual classroom? Between program students and the national norm group? Must you show the trend of language acquisition, or identify which parts of the program are contributing to student achievement? Will you need to tell how people feel about the program and why? Each of these information requirements leads to different statistical analysis procedures.

Generally, you will need to report statistics of two sorts: descriptive and inferential. You will need to describe both your instruments and something about the kinds of data they provided before making inferences about program effects. This preliminary data analysis will include instrument statistics such as inter-rater reliability for all constructed response measures, internal consistency and perhaps test-retest (pre and post) reliability for achievement tests, decision or classification

consistency for criterion-referenced measures, item difficulties, as well as group means and standard deviations.

In addition to descriptive statistics about your instruments, you will want to provide intercorrelations among dependent variables or measures as an indication of their similarities and differences. Finally, you will look at the data itself, test scores, questionnaire response tallies, observation tallies, writing sample scores or rankings, and interview responses in order to answer three questions:

What general patterns are emerging that can be examined with inferential statistics?

Are there missing data, unusual frequency distributions, small sample sizes, or restricted variance that will affect the interpretation of evaluation results?

Do the data to be used for making inferences meet the requirements for inferential statistics?

To examine patterns in the data, you will need to make tables, frequency distributions and even graphs so that you can "see" what is happening. Are student scores clustering at one end of the distribution? Do questionnaire responses cling to the safe "middle" response? Are some categories in the teacher observation schedule not being used at all? Do responses from parents cluster into two or three distinct categories?

The graphs and distributions you used to look at data, as well as the test statistics you computed for appropriate measures, will point out where the data may be doing "unusual" things. Are there missing data? Do you have pre and posttest scores for all or most of the students in the

program? Are distributions very restricted in range (small standard deviation) or are they so different from each other as to make comparisons among measures not useful? Your data display, or use of descriptive statistics programs such as "Breakdown," "Frequencies," "Crosstabs," and "Reliability" in Statistical Package for the Social Sciences (NIE, et al., McGraw Hill, 1975), will then help you check whether you have met the assumptions requisite for examining evaluation questions with inferential statistics.

In order to do significance testing, regression analysis, or analysis of variance, techniques used for comparing groups, your data should fit the following four requirements:

1. The data is reported on an interval scale. Scores, or responses are considered to differ in amount, degree, or kind along a continuous, linear scale. It is also assumed that growth along this scale is linear.
2. The scores or observations are independent. Factors influencing one score differ from those influencing another. This independence requirement will almost never be satisfied when data are gathered from intact classes; however, using classes as the unit of analysis may help to meet the assumption.
3. Students in the program have the same distribution of characteristics as the population from which they were sampled. Your sample is representative of the population of bilingual students to whom you will generalize.
4. The variances of any groups being compared on the same variable

are not significantly different from each other. These comparison groups have similar patterns of performance reflected in similar score distribution shapes and spreads.

If any of the above assumptions are violated, you will need to check the assumptions associated with particular statistical tests to see which can still be used. Inferential statistical analysis techniques are not discussed in this workshop; however, references for data analysis techniques appropriate for certain design questions are presented after the discussion of typical designs selected to answer that question.

Planning for Evaluation Design

Workshop B: Selecting Evaluation Designs

frequently asked in bilingual evaluation better than others. These recommended designs are ones that do not require control groups, random assignment, or comparisons with non-equivalent populations but still allow you to gather the comparative information required in many evaluations and to estimate no treatment expectations.

Although you will be presented with models of evaluation designs, both for general and special purpose bilingual evaluation concerns, no one design model will necessarily fit your needs. Designs are particular to the evaluation setting in which they are used, even though the principles of systematic data collection will remain the same. Thus, you will need to consider how you can adapt/employ certain representative designs to your program. This issue is considered in the section on Troubleshooting Design Problems, a guide for adapting designs to the conditions and events that operate uniquely in your program.

The final section of this workshop identifies designs that may be useful when special conditions exist in the bilingual program that allow you to use the more traditional "experimental designs." The required conditions and design requirements are summarized in Figure 2-7.

Introduction

This workshop is organized around five issues that you will need to consider when planning your evaluation design:

--What factors influence the selection of an evaluation design?

--How do I select an appropriate design?

--Are some designs more appropriate than others for bilingual evaluations?

--How can designs be adapted to the particular conditions of my program?

--Are there any other kinds of designs that might be useful in my evaluation?

The first question considers the factors influencing design selection. Before reviewing possible sampling, instrumentation, and data analysis strategies, you will need to consider the ways in which your bilingual program evaluation in both similar to and different from the other kinds of program evaluations that you have conducted. You should also try to anticipate the kinds of conditions affecting your program, such as a high student transiency, and other events, e.g., special state testing schedules, that may influence your evaluation design.

The second question, how do I select appropriate designs, is really a question about selection criteria. What are the variables that I need to examine in determining whether certain systematic data collection strategies are appropriate?

While it is always difficult to recommend certain designs over others, there are some evaluation designs that seem to address the questions most

Factors Influencing Design Selection

Before planning your evaluation design, you will want to consider the following questions:

How is your bilingual program like other programs in the school?

How is the program different?

What conditions and events can occur during the program that will affect the evaluation?

Similarities to other educational programs

In very many ways, the evaluation concerns of bilingual programs are similar to those of other programs in a school. It is often said that bilingual education is still an undefined treatment and that the variation in programs, teaching strategies and objectives make evaluation especially difficult. The same, however, may be said for reading or social studies instruction. We very often do not know what happens in classrooms. We make certain assumptions about the instructional program based on a course of study (that may not be followed), materials (that may be used in different ways or not at all) and test results (that have varying degrees of validity for particular programs). The problems of defining and documenting the treatment to be evaluated in bilingual evaluations, while more obvious, are analogous to those of evaluating other educational programs.

Similarly, the point is made that bilingual programs are often new or installed with less than hoped for enthusiasm. While this may be true, other innovative programs or mandated changes, such as Title IX and integration, may share these problems and provide guidance for the bilingual evaluator.

Other problems encountered by bilingual evaluators, subject mortality, lack of appropriate comparison groups, and narrowly-focused program guidelines are problems shared by other categorically-funded programs. Some of their solutions may have applicability to bilingual evaluation.

Most of the problems faced by bilingual evaluators may be found in other educational settings, but there are some factors special to bilingual programs that evaluators must consider in selecting designs.

Differences from other educational programs

Diversity. Although many educational programs differ in goals and methods, the within program diversity that characterizes many bilingual programs poses a special problem for the evaluator required to report comparative information. When teaching strategies and goals differ within a school, for example, which comparisons are appropriate to report? How similar must two classrooms or programs be to be comparable? Which elements of the programs being evaluated must be similar? When is diversity in treatment not a factor in evaluating program outcomes or effectiveness?

Assessment. Bilingual evaluators find themselves with the problem of assessing achievement of students whose native language is not English. When do test results reflect achievement and when are they a reflection of language proficiency? How can the two be separated? Achievement tests produced for English speakers are based on a curricula that may or may not be like those found in bilingual programs. The assumptions underlying test development need to be tested in bilingual settings before test validity is established.

When the issue of testing students in their native language is raised, the task becomes more complex yet. Basic questions about the kinds of skills that should be tested, the dialects that should be represented on the test, and test objectives have not been answered for native language achievement tests. What are reasonable achievement expectations of second-language learners? How do these differ according to native language type and proficiency level? Which assessment strategies cross language, culture, and instructional history?

Until these issues are solved, testing in native languages will remain controversial. There are no currently published tests that are widely applicable to bilingual evaluation. What, then, is the evaluator to do? You may wish to adapt instruments with input from local staff. You could administer locally developed or newly developed tests concurrent with tests of knew psychometric properties to determine the quality of the new tests while supplementing information gathered from the old.

When time and cost constraints make test development and validation impossible, you will need to gather descriptive data and use multiple measures of program outcomes. You can then examine how these differing data sources converge on certain conclusions and rule out competing plausible explanations for evaluation results.

Accountability. Because bilingual programs may operate under several kinds of guidelines, federal, state, and local, the evaluator may have several "masters" to serve. The implication this has, of course, is that accountability to these funding and regulatory agencies may become the overriding concern of the evaluation, [illegible]ing other program needs.

The bilingual evaluator has to be especially careful to select instruments that provide descriptive and program improvement information that will be of use to local staff. This information may very well come from the same instruments used to report accountability data; its interpretation and presentation may differ, however.

Conditions and events influencing bilingual programs. Bilingual programs often suffer a series of events that interfere with data collection. Figure 2-1 summarizes events in the three stages of bilingual evaluation that may hamper data collection efforts. You will notice that all of these events and conditions can be found in other educational programs; the difference is that you will find more of them in bilingual ones. What to do? Evaluation designs will have to be adapted to these program events. How you can use designs that are appropriate under these conditions is described in a later section, "Troubleshooting Design Problems."

Figure 2-1

Events in the Life of a Bilingual Program Evaluation Affecting Design

Before Program: Selecting a Design	During the Program: Adapting the Design		At the End of the Program	
EVENT	EVENT	REASONS	EVENT	REASONS
No comparison group	Attrition of program students	(May fluctuate seasonally)	Incomplete data received from instruments. Different instrument used for posttest. No change observed pre→post	
No instruments to measure language	New students populate the program. The additions may or may not resemble the current population.	(Busing may affect population)		
No local or other norms available			Year-end testing conflicts with required year-end state or other testing	
No demographic data available on the population	One treatment is lost or gained	(ie., new teacher arrives and changes approach; students are lost)		
Two or more language groups within the program	Students are transitioned from native language to English	(May occur at varying points during the year)	Posttest shows there is a different population at year-end from beginning	Mortality transfers into the program
No baseline data available for individuals or program				
No program goals; no criteria/standards for judging program effectiveness	More money and resources made available for the program and evaluation	(Funds arrive after program begins. New funding sources added)		
Tests prescribed by outside office or agency	Program begins during the school year. No pre-test data.	(Tests not used for selection. No appropriate pretest available)		
More than 1 level of achievement and/or language competence per grade				
No personnel available to measure students, develop tests or train	Students were mislabeled at beginning of program. Discovered later.	(Lack of placement instrument/ procedure. No clear exit criteria. Inexperienced staff		
Very small n of student in the bilingual program				

Selection Criteria

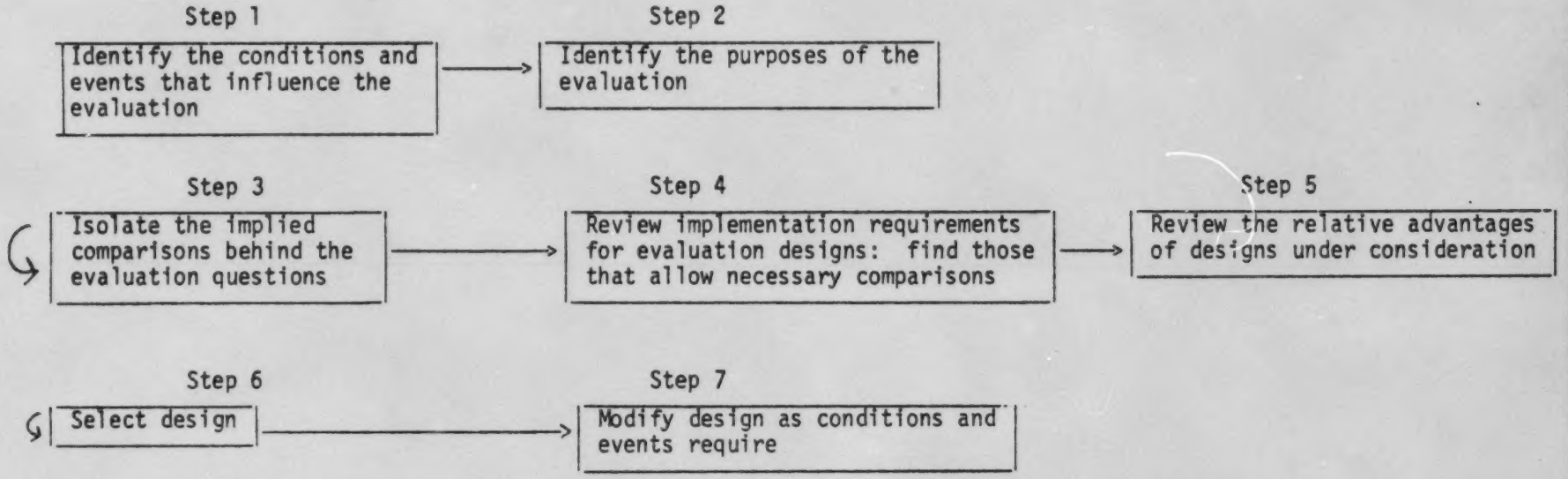
Selection criteria for screening potential evaluation designs include: the evaluation setting, the purpose of the evaluation, the comparative data required, requirements for implementing particular designs, and the relative advantages of using one type of design over another. The relationship between each of these factors and evaluation design is discussed in this section. A fuller description of the design models referred to is presented in the following and last sections of this workshop.

Evaluation setting

You will need to identify the variables constraining your particular evaluation before selecting a design. Discussion in the previous section about the similarities and differences between bilingual and other educational programs and of the special conditions affecting bilingual programs indicates how those constraints might be identified. (See Figure 2-2.)

Figure 2-2

Steps in Selecting Evaluation Designs


Step 1
Identify the conditions and events that influence the evaluation
Step 2
Identify the purposes of the evaluation
Step 3
Isolate the implied comparisons behind the evaluation questions
Step 4
Review implementation requirements for evaluation designs: find those that allow necessary comparisons
Step 5
Review the relative advantages of designs under consideration
Step 6
Select design
Step 7
Modify design as conditions and events require

Purpose of the Evaluation

The questions guiding your evaluation will determine the purposes for which your designs are selected. Some of these questions and their attendant purposes appear below:

Question	Purpose
What does the program look like?	Program documentation
Is it being implemented as planned?	Implementation evaluation
How can the program be improved?	Formative evaluation
Is the program promoting student achievement and other desired outcomes?	Formative or summative evaluation

The evaluation purpose and kinds of questions you will be asked to answer provide requirements for the kinds of data you will need, e.g., student achievement, attitude or self-concept, observations, questionnaire, interview, etc.

Comparative Data Required

Evaluation questions of the sort listed above often require the gathering of comparative information to make inferences about program effects or relative strengths and weaknesses. The need for comparative information does not imply that you must use controlled assignment techniques such as randomization or matching. Comparisons could include, for example,

- increases in fidelity to the implementation plan over time
- degree of conformity to the program description
- comparison of current program students with previous cohort groups
- changes in parent attitude or participation over time
- trends in student behavior, such as absence, tardiness, or office referrals

Figure 2-3 summarizes some typical camparisons that you might make in addressing five commonly asked questions in bilingual evaluations.

Design Requirements

Each design you select has different implementation requirements regarding sampling procedure, timing of tests, and data analysis. If you cannot meet these requirements, the design will be less effective in controlling competing plausible explanations for program results. The requirements for the four suggested bilingual designs and the nine special purpose data collection strategies are given with the discussion of each of the designs. You will find them in the following and last section of this workshop. Figure 2-4 summarizes the requirements common to all designs and provides page references for particular ones.

Relative Advantages of Designs

Designs are used to make evaluation results more credible by ruling out plausible, competing explanations for results. Some designs rule out certain competing explanations, such as those posed by the sample, while others appear to be better at limiting alternative explanations due to instrumentation. Some threats to valid evaluation interpretations, e.g., mortality, improper test administration, and invalid instrumentation, cannot be countered very well at all. You will need to decide which alternative interpretations pose the most threat to the credibility of your particular evaluation, and then consider the advantages of designs relative to this decision. Figure 2-5 displays relative advantages of the thirteen designs described in this workshop.

Figure 2-3

IMPLIED COMPARISONS

Evaluation Question	How did students in the program do?	Did the program improve self concept and cultural awareness?	Were parents involved?
Implied comparisons	Compared to similar students Compared to English monolinguals Compared to a national norm Compared to a prespecified standard of excellence Compared to their past performance	Compared to what it was? Compared to ideal standards Compared to a non-program group's responses Compared to a scale established on a published instrument	Compared to the past Compared to non-program parents Compared to other bilingual program parents Compared to the PTA, PAC, booster club

Evaluation Question	Was the program implemented as planned?	How can the program be improved?
Implied comparisons	Compared to the course of study Compared to the description in the funding proposal Compared to what teachers said they were doing Compared to community expectations Compared to other bilingual programs Compared to other academic courses	Compared to the association of different parts Comparing the relative achievement produced by different parts of the program Comparing the affective outcomes produced by different parts of the program Comparing selected groups' (students, teachers, parents) judgment of how well they liked different parts of the program Comparing the relative implementation effort expended on different program parts

Figure 2-4

Requirements of Typical Design Models:
a Reference

Model	Comparative Data Obtained	Text Reference For Design Requirements
1. Time Series/Longitudinal	Within Group	96-97
2. Multiple Time Series	Within Group	96-97
3. Exposure to Treatment	Program to itself	101-105
4. Minimal Competency	Prespecified Achievement Standard	91-92
5. Normed Pretest, Normed Posttest	National Norm Group	82-85
6. Normed Pretest-Local Pretest-Posttest	National Norm Group	86-88
7. Non-equivalent Control Group	Non-equivalent Groups	70-72
8. Regression Discontinuity	Non-equivalent Groups	73-75
9. Regression Projection	Non-equivalent Groups	76-77
10. Pretest-Posttest Control Group	Equivalent Group	58-60
11. Posttest Only Control Group	Equivalent Group	61-63
12. Single Cohort Comparison	Equivalent Group	64
13. Multiple Cohort Comparison	Equivalent Group	65-66

Requirements for all models

1. Theory/knowledge of developmental rate of skill being examined
2. Evidence that the program existed
3. Description of program and sample "mortality" with reasons if possible
4. Documentation of activities that may produce competing explanations of results
5. Instruments - valid and reliable
6. Explicit criteria by which program "success" will be judged

Figure 2-5

RELATIVE ADVANTAGES OF TYPICAL EVALUATION DESIGNS

Threats to Validity Countered by Design

	Sample			Instrumentation			Extraneous	
	Selection	Maturity	Mortality	Validity	Reliability	Admin. Procedures	Concurrent Programs	Hawthor Effect
Pre-Post Control	X	X					X	
Post Control	X	X	X (somewhat)				X	
Single Cohort	X						?	
Multiple Cohort	X						?	
Non-equivalent Control	X							
Regress. Discontinuity					X			
Regress. Projection					X			
Norm Pre-Post					X			
Norm Pre/Local Pre-Post					X			
Minimum Competency		X		X				
Time Series	X	X					X	
Multiple Time Series	X	X					X	
Exposure to Treatment		X					X	X

Recommended Designs

Introduction

Design models are symbolic representations of actual strategies for carrying out research and evaluations. They symbolize the kinds of comparisons made, assignment conditions (random or intact groups) used, and number of measures or observations taken. Most evaluations incorporate several kinds of designs, depending upon the questions guiding the evaluation. Certainly it is very unlikely that an evaluation design appropriate for your program would look exactly like any one of the models presented here.

The purpose of this section is to provide you some exemplary designs suitable for gathering information related to the major kinds of questions that you would most likely be addressing in bilingual evaluation. You may wish to refer to some of these models, focusing on their strengths and weaknesses as they relate to your particular evaluation concerns. In this way you can adapt, combine and otherwise apply the design concepts they embody when you plan your own evaluation design.

The designs presented in this and in the final sectin of the workshop are organized into two groups: recommended designs and special purpose designs. The designs themselves are grouped by the kind of information they allow you to gather: historical, treatment specification, criterion-referenced, norm-referenced and so on. Groups of designs are discussed in terms of the general evaluation purposes they serve, typical questions answered, and the kinds of comparative information they provide. For

each of the design groups, we provide specific design models, identified by their experimental-design designation (Campbell and Stanley, 1963; Cook and Campbell, 1979). The design models are pictured and described in terms of implementation requirements, with advantages and disadvantages of each model given. This discussion ends with data analysis references for each of the models.

Notation

The presentation of design models is simplified with the adoption of some simple notation. This scheme, adapted from Fischer, was employed by Campbell and Stanley in a book that has become the standard work on design of Field Studies (Campbell and Stanley, 1963). As a result, the notation is now almost a universally accepted way to depict research designs.

Each experimental situation is summarized in a diagram. The diagram is read from left to right, with the understanding that the time sequence proceeds in the same manner. The following symbols are employed in the various diagrams:

R signifies that subjects have been randomly assigned to a particular group

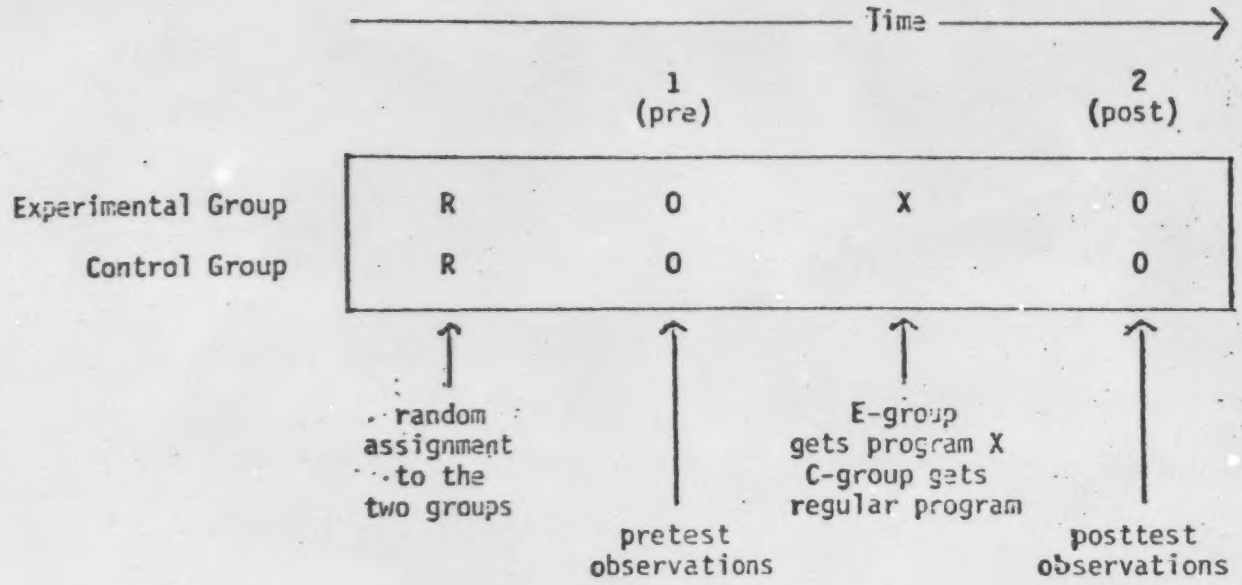
O designates a planned observation, that is, a measurement such as a test, interview, case study, performance test, etc.

X designates an occurrence of the experimental treatment or the special program

----- a broken horizontal line separating two groups indicates that the groups are non-equivalent, e.g., they were not formed by random assignment.

The experimental or program group is normally described on the top line and the control group (if one exists) is presented below it.

For example, a program in which students were randomly assigned to either experimental or control groups and in which both pretesting and posttesting occurred would be diagrammed as follows:



In order to make the diagrams clear and easy to read, one important event is omitted. No symbol is included to indicate what happens to the control group during the time that the experimental group is receiving the special treatment. It is understood that the control group is engaging in normal educational activities during the time that the experimental group is receiving the experimental program.

Recommended Designs for Bilingual Programs

The four design types suggested for bilingual programs are summarized in Figure 2-6. Each of these types was selected to address a different, but frequently encountered, bilingual evaluation purpose. The designs are presented in order of utility, from general to specific. The most general purpose designs are the time series/longitudinal for investigating program outcomes. The exposure to treatment design is widely applicable for formative evaluation purposes. Accountability designs, for reporting information about student achievement either in terms of local program objectives or national norms (when required) are presented also.

The discussion of each of the design types begins with the kinds of questions most often asked for which the design is used, and why this question is important. Sampling and instrumentation requirements are also discussed. The general discussion ends with the threats to validity not countered by the design types. This discussion is followed by specific examples of the design types, the implementation requirement of the design and data analysis references.

Figure 2-6

Recommended Designs for Bilingual Programs

Evaluation Purpose	Design	Implied Comparison	Requirements of the Design
All purpose design to provide documentation of program development, implementation, and student achievement over time	Time Series/ Longitudinal	Within group or age/grade cohorts	1. Way to trace/document departure of students from bilingual or comparison groups to explain "mortality" 2. Tests/instruments that provide comparable data over time, e.g., same test with many levels and alternate forms, observation scales that remain the same across the period of the program evaluation or highly correlated instruments.
Identifying program strengths and weaknesses in relation to student achievement	Exposure to Treatment	Parts of the program with each other	1. Identified theory of what variab contribute to bilingual student achievement 2. Achievement instruments that measure program goals 3. Data gathering techniques that identify qualities of program implementation such as time on task, amount of curriculum covered, etc.
Accountability to school or district administration	Minimal competency	To prespecified minimal performance standard and/or to students in district outside the bilingual program	1. Basic skill/minimal competencies that bilingual students are responsible for completing 2. Reliable and valid minimal competency test 3. Defensible minimal competency standard of excellence
Reporting required norm-referenced test data for Title VII programs	Normed pretest-Local pre-posttest	National norm group Other students taking local pre and posttest	1. Norm-referenced instrument appro priate for measuring bilingual program goals. 2. Reliable and valid local criteri referenced test that matches program goals 3. (Optional) Local norms for CRT

Historical

Evaluation purpose: All

Typical Question: How are outcomes of program participants changing over time?

Comparison: Within group

Why This Is Important: Some important changes which occur as a result of the bilingual program may be cumulative in nature. Students who show little gain in the first two years may suddenly bloom in the third. To condemn such a program based on the outcomes from a single, early year alone is unwise. By taking repeated measures over a long period, it is possible to observe change without resorting to comparisons with dissimilar groups. While longitudinal designs do have their problems, they are often able to uncover variables affecting outcome which are the result of the continuing program and not just isolated phenomena.

Who serves as the sample?

Sometimes no equivalent groups exist, school turnover makes cohort comparisons and single group longitudinal studies impossible, and adequate instruments for norm group comparisons cannot be found. This still leaves a very powerful comparison for the evaluator, changes over time. While within group comparisons don't tell you how well students would have done without the program (the focus of the comparisons in the previous models) for most bilingual programs this is not even an issue. There is a program. The program is there because it is needed. What is left to investigate is, what happens to the students as they participate in the program? Are they growing slowly? Quickly? Evenly? In a disjointed fashion? How? This is perhaps the most valuable information of all because it is both addressing the proper question, what happens to students in the program, as well as providing information about the learning process itself.

This design utilizes the performance of the experimental group itself as the measure of program impact. A time series design like the cohort designs attempts to establish a recognizable pattern of achievement prior to the implementation of the program, against which post-program achievement will be compared.

What data should be gathered?

The key to within group comparisons lies again in instrumentation. However, taking many observations both before, during, and after instruction can compensate in part for some deficiencies in instrument quality. The timing of observations for within group comparisons is important. As many observations as are needed, careful planning is needed to prevent testing from intruding on instruction.

To implement a time series design, an evaluator needs to gather data on the students before the program is begun. In many schools, achievement data from prior years can be obtained and can serve as the first observations. Test administration must be spaced far enough apart to control for testing effects. Once the treatment is initiated, the pattern of periodic testing should continue while the program is carried out.

Identifiable changes in the pattern of achievement after the initiation of the program will then be a measure of the program's effect, but these patterns must be examined carefully. It is not unusual for rapid initial gains to decline over time. As a result, the evaluator should not base his or her conclusions on only one or two post treatment assessment. Other competing explanations for the results may exist, and the evaluator who is considering a time-series design would be wise to familiarize himself or herself with some of the material that has been written about interpreting the results of such an analysis. For this purpose, we suggest Cook and Campbell (1979).

Since programs, as well as students, change over time, you will also need to collect information about program activities, materials, objectives and such and note how they have evolved with the passing of time.

What are the threats to validity encountered in the design?

Threats to validity in making inferences from time series designs stem from two principal sources: instrumentation and loss of students (subject mortality).

1. The model can be greatly affected by differential loss of participants over time. This subject mortality (loss of students) in the latter grades may be a possible explanation for observed changes over time.

2. You must be able to use tests which control or accommodate maturation.

3. If standardized norm-referenced tests are employed, they are likely to be more sensitive to changes between earlier grades than between later grade levels.

4. You may need a comparison group of the same age but not in the program to control for maturation. This means you are looking at more than one group in the same grade or level in the bilingual program. If there is only one bilingual group, cross-sectional analysis could mean collecting data for non-program people.

5. If the same test is used at multiple points in time, then the instrument should have alternative forms to reduce "practice" effects.

6. Outcomes may be due to school and unplanned program changes. Therefore, you must document non-program school changes as well as changes within the program.

Model 1: Time Series Design

The time series design is illustrated by the following diagrams:

	Time						
	1	2	3		4	5	6
Experimental Group	0	0	0	x	0	0	0

Model 2: Multiple Time Series

	Time						
	1	2	3		4	5	6
Experimental Group	0	0	0	X	0	0	0
Cohort Group	0	0	0		0	0	0

Description

1. The most important part of this design is scheduling the testing or observations. Several must be taken before the instructional intervention, and several after.

2. The observations do not all have to be taken during one school year.

Advantages of Model #1

1. This design is "easier" to use because it does not require a control group.
2. By focusing on long term patterns of growth, this design may give a more accurate reflection of the pattern of student achievement in school and program impact.
3. Results are easy to communicate.

Disadvantages of Model #1

1. Because there are no comparison groups, it is difficult to rule out competing interpretations to the results. One lacks the ability to say how one program compares with other alternatives,

or whether student achievement is a reflection of selection-maturation or program.

2. It is difficult to rule out maturation, selection, mortality and instrumentation biases from the results.
3. One cannot always identify a particular moment when a program is officially begun. In the schools, implementation is often more gradual than distinct; thus, which measures are "pre" and which are "post" is not always clear.
4. It is impossible to go back in time. Thus, if archival information is not available from prior years, it will be difficult to use a time-series design.
5. The additional data collection required by this design will require extra time and resources.
6. The amount of testing required may influence student achievement by cueing them into the goals of the program. Gains made under these conditions may not be replicable in designs requiring less testing.

Treatment Specification

Evaluation Purpose: Formative/Identify Program Strengths

Typical Question: Which treatments within a program produce different (or better) results?

Comparison: Program to itself

Why This Is Important: Bilingual education is an evolving entity. Teachers are experimenting to discover which methods work best with which students as they add to their native languages new skills in English. Therefore, it is especially important to enhance the state of the art by documenting the various treatment procedures teachers, or groups of professionals, use in their daily routines. These can then be compared for their effectiveness, and the most useful approaches adopted as a matter of policy or recommendation.

Here student outcomes are used to look at the relationship between student outcomes and the process of instruction.

Within a single school there may be as many interpretations of "bilingual education" as there are teachers. For this reason it is crucial to document within-program differences in approach, materials, and method. Outcomes which appear unimpressive may suddenly show stunningly positive results when examined from this perspective. In addition, new information which can improve the state of the art may be found. For example, if it is shown that the teacher using mainly English to instruct her class of fourth-graders has higher performance means than a teacher using a different method, then one might conclude that for that type of student, a high proportion of English leads to quicker transition. Then, when the bilingual program is being planned for the coming year, teachers may use this knowledge to provide the most appropriate service to the pupils.

If you can identify the parts of the program promoting student achievement or other desired outcomes and parts that are not working well, you then have guidelines for program improvement. The formative evaluation focus of treatment specification-type designs is an

especially useful perspective in bilingual evaluation. The emphasis shifts to "How can this program be improved?" from "Is this program any good?"

Who serves as the sample?

Since you will be making comparisons between quantities of program and student outcomes, you will need to sample both variables in the program as well as students. Such variables could include teacher experience, minutes of instruction in the native language, time on task, amount of time spent in small-group work, amount of time in reading instruction, number of workbooks completed, etc.

What data should be gathered?

The key to the effective use of the treatment exposure design to describe the relationship of program to student outcomes is finding an effective index of program exposure. Some suggested variables include amount of time spent in bilingual instruction, amount of time on task spent by students, amount of material covered relevant to instructional goals, instructional strategies related to program goals and so on. Planning for this design involves careful selection of the pre-posttest instruments and careful specification of those aspects of the bilingual program that teachers think "exposure" to will result in boosting achievement.

Assumptions of the design

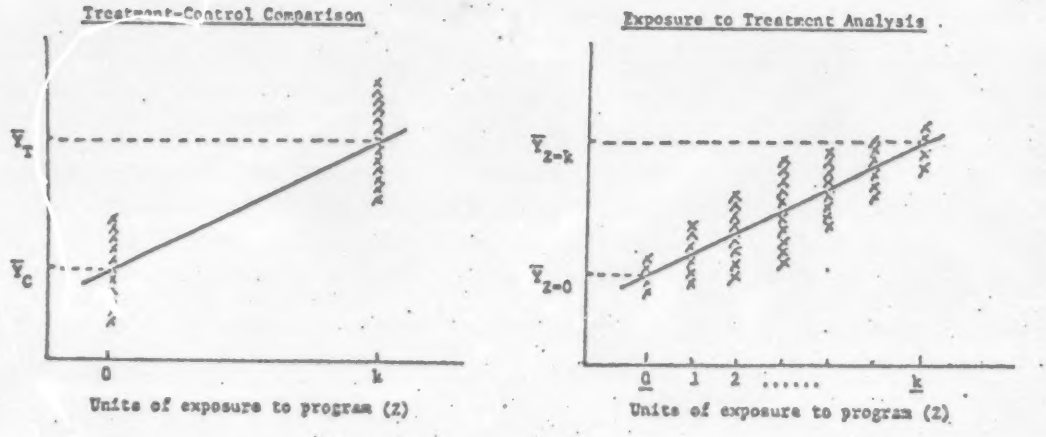
1. Student achievement and growth without the program is linear.
2. Impact of program effects are linear over period of the program.
3. There will be variation among classrooms/units sampled in the degree of exposure to the treatment.

What are the Threats to Validity Encountered in this Design?

1. Instrument validity poses the most serious threats: measures should assess program aspects that are related to student outcomes.
2. The strength of the theories of learning and instruction that you use to identify possible program-student outcome relationships is also crucial for valid inferences to emerge from this design.

Model 3: Exposure to Treatment

FIGURE 2-3



Description

1. The design requires three pieces of information to be available, a measure of students' performance prior to the beginning of the program, an index of exposure to the program (compiled from implementation data), and a measure of students' performance after the program.

2. In order to interpret the relationship between outcome and exposure to treatment, it is necessary that the pretest be a completely adequate summary of what the students knew about the material to be taught before actually being exposed to the material.

3. When the portion of the end of the program achievement that can be predicted from the pretest is removed, the remainder of the posttest gain to the extent it covaries with amount of exposure to treatment may

legitimately be attributed to the amount of treatment students received. While the pretest does not have to be precisely the same test used for measuring program outcomes, it must be an adequate summary of students' mastery of the relevant content prior to instruction. Thus, even though the goal of the design may be to describe the impact of the bilingual program on the students' norm-referenced achievement, the pretest need not be a norm-referenced test.

4. The logic behind the design is of two sorts. First, if students are tested before and after the program, and data reflecting the amount of instruction received during the interval between testing occasions is also collected, you can relate amounts of gain shown by students to the amount of instruction they received. By this reasoning the analysis of the data divides the achievement gain in test scores into two parts, one that is predictable on the basis of knowledge of the students' pretest scores alone, i.e., the part of the gain that would be expected even if the program was not implemented, and another part that is apparently related to exposure to the program. If this latter part of the gain is substantial, it is evident that the program increased students' scores beyond the level we would have predicted knowing only how the students performed before entering the program. The analysis plan does not imply that all students' scores were improved by the program; rather, it suggests that the more program-related instruction students receive, the higher their scores will be.

The strategy for estimating exposure to treatment involves estimating the amount of achievement gain that is associated with increments in exposures to instruction, other things being equal. In the case of outcome

scores derived from a nationally normed test, this means that the analysis will provide an indication of the extent to which students standing in comparison to national achievement levels, could be improved by participation in the bilingual program. We use the words "could be" because the analysis does not assume that all students have been exposed to influential amounts of bilingual instruction. By imagining a variable that measures how much bilingual instruction students receive, we can show how the strategy relates to Models 7-13. In Models 7-13, one attempts to determine whether being in the treatment group is associated with higher outcome scores, on the average, than are obtained in the control group. This is equivalent to the extreme cases in the program exposure model, where each student receives either no bilingual instruction, i.e., a score of zero on the imaginary variable, or some presumably adequate amount of instruction.

Figure 2-3 portrays the analogy between the traditional Treatment-Control analysis and the program exposure analysis. In the left hand drawing, the zero for the control group and the "k" for the Treatment group, on the vertical axis show the amount of gain exhibited by the two groups. The right graph depicts the single group case where more fine-grained implementation data are available about the amount of the treatment each student in the group received. If the amount of the treatment ranges from zero units to "k" units, we can see that the average amount of gain for those students in the group with zero and "k" units of exposure to treatment is the same as for the two groups in the left graph. This equivalence is further demonstrated by the fact that we can fit the same "regression" line to the data in both graphs.

Advantages of Model 3

1. It can be used when only one group is available.
2. It can identify "effective" program components, when effective is defined as gains in achievement.

Disadvantages of Model 3

1. The exposure to treatment design depends upon assumptions that may be unrealistic. The assumption that without program intervention student growth follows a straight line, an assumption underlying all regression models, is not necessarily true for all kinds of learning.
2. Finally, this design does not support any comparative statements about the relative merit of your program and some others.

Data Analysis References

Capell, F., Sisson, B., & Winters, L. Final Reports Basic Skills Learning Evaluations. Center for the Study of Evaluation, September 1979.

Kerlinger, F., & Redhauzer, E. Multiple Regression in Behavioral Research. Holt-Rinehart & Winston, New York, 1973.

Criterion-Referenced

Purpose: Accountability/Program Improvement

Typical Question: How well are students doing compared to a prespecified standard?

Comparison: To a prespecified standard or criterion

Why This is Important: With the emphasis recently on basics and school accountability, observers want to be assured that that students learn specific skills. Criterion-referenced tests have been developed in order to test objectives of instruction. These tests suggest that mastery of specified skills fulfill educational objectives. In answering this question, students' performance on an instrument relevant to their curriculum is compared with a level of performance deemed acceptable. The determination of that level is an important issue. Since any standard is arbitrary and arguable, care must be taken to select one that can be disputed by the fewest number of critics.

Answering this question may be the only acceptable evaluation when there are no other appropriate comparison groups available. This can occur when:

... Change in planned program or comparison groups makes them dissimilar (attrition or change in program).

... Norm groups for prepared tests are dissimilar from program students, or no norms are supplied.

... There are no normed tests in the language used in the bilingual program.

... The program teaches content (e.g., cultural awareness) which is important but not measured by prepared tests.

Who serves as the sample?

You may wish to gather data about typical performance of students not enrolled in the bilingual program as well as bilingual students for comparison purposes.

What kinds of data could be gathered?

The pre-specified criterion design requires that you select some criterion of adequate performance as your pre-posttest instrument. This could include a criterion-referenced test, a curricular embedded test, work samples or performance tests of some sort.

In order to satisfy funding agency requirements, you may have to gather local norms on your criterion-referenced instrument and/or equate it with a

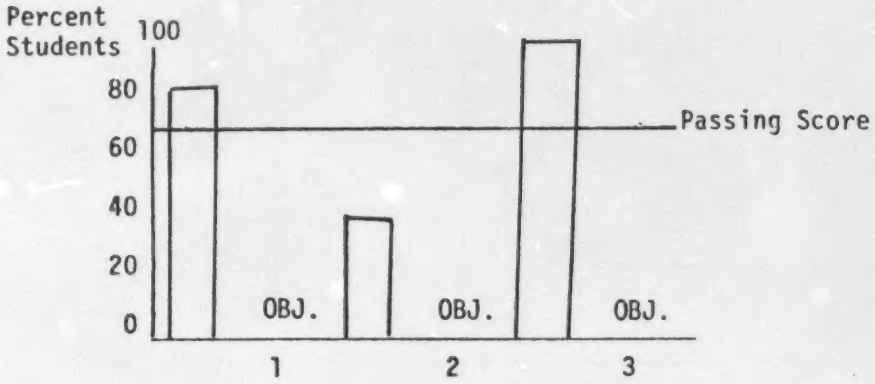
norm-referenced test as in RMC model A. Since criterion referenced tests are designed to reflect objectives upon which local school staff put emphasis, you will need evidence that 1) these objectives are part of the program, 2) teachers are indeed teaching to these objectives and other kinds of appropriate program implementation and description information, and that these objectives coincide with those of the funding source.

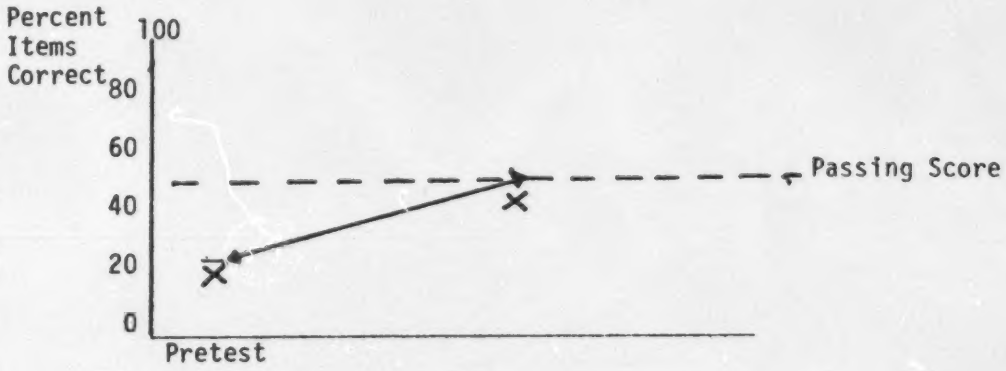
What are the threats to validity encountered in the design?

a) All standards are arbitrary and arguable, and must be set in a manner acceptable to the community, funding agencies and others.

b) Results are meaningful only if test difficulty is appropriate to the level of the students being tested.

Model 4: Minimal competency design





Description

1. The minimal competency design is simply one type of criterion referenced comparison that you may choose to make. For minimal competency comparisons, you may select those objectives and performance standards set by the school district for compliance with minimal competency legislation.

2. Students in the bilingual program are pretested and posttested with the district established minimal competency test.

3. Since minimal competency testing is part of an early-detection, instructional intervention model of schooling, you must be sure that students in the bilingual program are receiving instruction on objectives tested by the minimal competency examination.

4. Comparisons are made both between bilingual pre and posttest scores and between the number of bilingual students able to achieve the district minimal competency passing score on two occasions.

Advantages of Model 4

1. Criterion referenced comparisons allow you to select standards for program success that may be more appropriate for innovative programs. Such standards could be in terms of minimal competency, parent participation levels, increases in student attendance, progress on curricular embedded tests and so on.

2. Criterion referenced comparison often provide important information for instructional decision making and program improvement.

Disadvantages of Model 4

1. The selection of the criterion for adequate performance is extremely difficult and not always agreed upon by different stakeholders in the evaluation.

2. Setting of performance standards for bilingual students is riddled with difficulties as to what is to be expected and how much difference between monolingual and bilingual student performance is to be accepted.

Data Analysis References:

Popham, W. J. Criterion Referenced Measurement. Englewood Cliffs, New Jersey; Prentice Hall, 1968

Linn, R. L. and Slinde, T. A. The determination of significance of change between pre and posttesting periods. Review of Educational Research, 1977, 47(1), 121-150

Norm Referenced Designs

Evaluation purpose: Accountability

Typical Question: How well are students in the program doing compared to a norm group?

Comparison: National or local norm group

Why is this important? This design is often used because it meets the criteria of providing comparative data that many funding agencies require.

Norm-group comparisons are hardly new in education. Almost all standardized achievement tests are "norm-referenced." This means that the tests have been administered to a large number of students representing the national student population (the norm group), and the performance of this group of students is used to standardize the scores on the test. From that point on, every other student who takes the test receives a score which compares his or her performance to the performance of the original norming group. Of course, test norms are updated over time by administering the test to newer groups of students and adjusting the scoring scales to reflect any changes in general student performance. Thus, a norm group comparison is one in which a student's performance is compared with the performance of a larger group, often a national, but sometimes a local, sample.

There are some serious questions about the validity of norm-group comparisons when one is considering very specialized populations such as Limited-English Speaking or Non-English Speaking students. They will be discussed below. In spite of these questions, norm group comparison are widely used both in Title I and Title VII evaluations. They have received such wide usage for two reasons:

1. There is a lengthy tradition of norm-referenced testing in this country. This type of assessment has wide acceptance and has served educators well in making selection decisions for many years.
2. Legislation often requires that evaluations answer the question "How would the performance of the students differ if they had not been in the program at all?" They require an estimate of what the relative performance of students on some normed instrument would have been without the program. Such measurements of "no-treatment expectations" are often made in terms of national norms when no other comparisons are possible. (Norm-group comparisons, however, do have serious drawbacks.)

The accuracy of a normed instrument for assessing performance of a particular group is directly related to the nature of the group. Because the test was originally "normed" on a general population, its measurement scale is only accurate for groups similar in composition to that population. This is a serious problem for bilingual educators, because, almost without exception, the most widely accepted norm-referenced tests were normed on an English-speaking population only. It is not known whether these tests accurately measure the relative achievement of Limited English Speaking and Non-English Speaking students or not.

Although there is every reason to believe that the scores LES/NES students obtain on such tests reflect something about their performance relative to the national group, there is no way to know exactly what these scores mean, nor whether they are at all accurate as a measure of

changing student performance over time. The assumption that bilingual students' growth patterns in a pretest-posttest comparison will approximate the growth patterns of the norming groups is highly suspect.

Notwithstanding these objections and others, norm group comparisons may be the only tool available to the bilingual evaluator. Over time, they may prove to have been valid measures of educational progress. For the moment, there is no question that they will continue to be widely used.

Who serves as the sample?

Students enrolled in the bilingual program serve as the sample. In programs where students have different linguistic backgrounds, or levels of language proficiency, you may want to consider breaking your sample into several sub-samples based on one or more of the following:

a) achievement level in one, two or more academic areas

b) length of time in the bilingual program

c) length of time in United States

d) attendance groups (e.g., students attending regularly, students with 15 absences, etc.)

What kinds of data should be gathered?

This design requires the use of a norm-referenced pretest and either a norm or criterion-referenced posttest. (In selecting a published test, it is extremely important to choose an instrument relevant to your own curriculum.) In order to assure that the norm groups of published tests and your population are comparable, it is best to collect demographic and recent achievement data on your own students. If you are unsure about characteristics of the norm group of the test you are using, first check the test manual, then call or write the publisher for details. You may also want to collect local normative data. This should be collected from students similar to those in your program and well in advance of program assessment.

Program implementation data, program input data, and context data may also be gathered in these designs.

What are the threats to validity encountered in the design?

Threats to validity encountered in the design emanate from two sources: instrumentation quality and relevance and appropriateness of the comparison group.

1. The national norm group differs drastically in linguistic and socio-economic status composition from most bilingual program participants.

2. The assumption that bilingual students' growth patterns from pre-to posttest will approximate those for the norm group are unsubstantiated.

3. The assumption behind normative comparisons is that if there is no pretest to posttest percentile gain for the bilingual group, then the treatment was not effective. There is some evidence that "disadvantaged students" in the absence of appropriate educational interventions, tend to fall behind their peers. In this case, a program that achieved no gain from pre to posttest, but also showed no loss in percentile ranking, might be considered effective.

4. If students selected for the program were chosen because their scores on a norm-referenced test were extremely low, gain might be explained by regression toward the mean (the tendency for extreme scores to be less stable) rather than program effects.

There are two designs presented here. Both are adapted from the models developed by the RMC Corporation for Title I evaluations. Further information about how these models are implemented and the data analyzed

can be obtained from the RMC publications listed at the end of this text. The two models are:

Model 5: Normed pretest-Normed posttest Design

Model 6: Normed pretest-Local pretest and Local Posttest Design

Model 5--Normed pretest, normed posttest design*

Model A is the first of the norm-group comparison designs. All norm-group models are essentially pretest-posttest comparison without a local control or comparison groups. The national norming group of the test is the hidden comparison group. That enters into the design, indirectly, in terms of the scoring scheme of the test itself.

If you were to diagram the normed-pretest-normed posttest it might look like this:

	$Time_1$ (pre)		$Time_2$ (post)
Experimental Group	O	X	O
(Norm Group)	(O)		(O)

*This model was developed by RMC Corporation and is often referred to as RMC Model A1.

Description

Certain cautions should be observed in implementing this design.

1. The pretest and posttest administrations should occur within two weeks of the midpoint of the empirical norming dates provided for the test itself. As noted previously, there is some debate as to the validity of normed instruments on sub-populations, such as LES/NES students. Thus, it is quite important that other procedures be followed as accurately as possible to insure the results have the greatest validity. Normed scores provided by test publishers are only valid for the times of the year in which the tests were actually normed. This varies from test to test. The evaluator will have to do some checking to find an appropriate test that has both fall and spring norming scales.

2. Schedule the pretest and posttest over as long an interval as possible. A good rule of thumb is that it should be at least 2/3 of the school year.

3. Scores reported on normed tests are not necessarily based on equal interval scales. Thus they cannot always be added and subtracted in the form they are reported. Percentile scores, for example, are not an equal interval measure. In order to analyze the data it is necessary to use test publisher's standard scores with equal interval scores such as Normal Curve Equivalent Scores (NCE scores) and scale scores. This procedure is described in Talmadge and Wood (1976).

4. If students are assigned to the program based on their performance on the norm-referenced test, then they should be pretested again after they have entered the program. This is to guard against regression

effects on low scoring, because the bilingual students are considered a sample from the same population as the norm group.

5. The same level of test should be used for both pretest and posttest.

6. Once you have selected and scheduled the norm-referenced test, and converted scores into standard or scale scores, compute the pre and posttest means and standard deviations.

7. Compute the pre and posttest correlations.

8. Look up the percentile or normal curve equivalent score for the pretest standard score. Do the same for the posttest standard score.

9. Look up the standard or scale score that corresponds to the pretest percentile rank in the posttest norm table.

10. If the actual posttest mean standard or scale score is larger than the mean scale score obtained in Step 9 above, test the statistical significance of the difference between the two.

Advantages of Model #5

1. This design can be used when no comparison groups are available and when you cannot or do not wish to pretest and systematically assign students to the program based on their pretest score.
2. This design utilizes nationally normed tests, and such information is often required by the funding agency.
3. The results are easy to interpret and present.
4. The analysis does not require sophisticated data processing capabilities.

Disadvantages of Model #5

1. There is some disagreement about the validity of the available norm-referenced instruments for assessment of bilingual populations. Thus, the results of the analysis are subject to instrumentation bias.

2. Norm-group comparisons may not provide much useful information for program improvement. This suggests making a normed pretest-normed posttest analysis part of a multiple-design evaluation strategy.

Data Analysis References:

Horst, D. G. Guide to measuring project impact on student achievement. RMC Research Corporation, Los Altos, California, 1975. Government Printing Office Stock Number 017-080-01460.

Tallmadge, G. K. and Wood, C. User's Guide: ESEA Title I Evaluation and Reporting System. RMC Research Corporation, Mountain View, California, 1976.

Tallmadge, G. K. and Horst, D. A procedural guide for validating achievement gains in educational projects. RMC Research Corporation, Mountain View, California, 1976. Government Printing Office Stock Number 017-080-01516-1.

Model 6 : Normed Pretest, Local Pre-posttest Design*

The normed pretest, local pre-posttest design is quite similar to model #5, the normed pretest, normed posttest design. The difference is that model 6 uses only one administration of a nationally normed test as a benchmark for calibrating scores on another test (one which assesses the same skills, but which may provide more useful information to the local program staff). Model 6 would be diagrammed as follows:

	Time	
	1 (Pre)	2 (Post)
Experimental Group	O_1 Norm X O_2 Other	O
(Norm Group)	(O)	

Description

When a non-normed instrument seems most appropriate for evaluating a particular program, it may be possible to "associate" it with a normed instrument and meet requirements for norm group comparison data. This is done in the following manner.

1. Two tests are selected that measure similar skill areas, one of which has been normed on a national population.

* This is also known as RMC A2 design.

2. Both tests are administered as pretests to all program participants within two weeks of the normed instrument's empirical norming date.
3. At the conclusion of the program the non-normed test is administered to program participants as the only posttest.
4. The average Normal Curve Equivalent (NCE) score on the normed pretest becomes the no-treatment expected score for program participants at the end of the program.
5. The students' scores on the non-normed posttests are converted to NCE using the relationship between the two pretests. This procedure is described in Tallmadge and Wood (1976).
6. The difference between the converted posttest score and the no-treatment expected score is the measure of the effectiveness of the program.

The ability to use a non-normed instrument to measure student growth is an important difference between Model 5 and Model 6. This allows the evaluator a much wider choice in test selection. Careful test selection can increase the meaningfulness of the results because better and more reliable bilingual assessment devices are being developed by a number of organizations. Thus, this analysis depends heavily on the validity and reliability of the non-normed instrument. The burden of test compatibility is shifted from the norm referenced test, about which a great deal is known, to the non-normed test, which might be of dubious quality. While there are difficulties interpreting the meaning of norm-referenced test scores with LES/NES students, there is no question about the general reliability of the instruments. The use of non-normed tests

introduces a new variable into the analysis. The evaluator who is going to use Model 9 should select the non-normed test with great care, examining the data on the test's validity and reliability in detail.

Advantages of Model #6

1. This model has all the advantages of Model 8 and, in addition, it allows the evaluator greater flexibility in selecting appropriate instruments to measure program impact.

Disadvantages of Model # 6

1. More time and effort is required to administer two pretests.
2. Non-normed instruments must be selected with great care. This design depends almost entirely on the validity and reliability of the non-normed test.

Data Analysis References

Horst, D. G. Guide to measuring project impact on student achievement. RMC Research Corporation, Los Altos, California, 1975. Government Printing Office Stock Number 017-080-01460.

Tallmadge, G. K., & Wood, C. User's guide: ESEA Title I evaluation and reporting system. RMC Research Corporation, Mountain View, California, 1976.

Tallmadge, G. K., & Horst, D. A procedural guide for validating achievement gains in educational projects. RMC Research Corporation, Mountain View, California, 1976. Government Printing Office Stock Number 017-080-01516-1.

Troubleshooting Design Problems

You will need to adapt the recommended and special purpose designs as program conditions and events change. Although it is impossible to anticipate all of the ways in which your evaluation setting may differ from the ideal, the conditions and events summarized in Figure 2-1 earlier will provide broad indications of trouble spots.

This section presents a Troubleshooter's Guide to dealing with the conditions and events that can alter your evaluation design. The Guide has recommendations for dealing with events that arise while:

--selecting the design

--implementing the design

--gathering data

You may wish to remove this section for handy reference.

CONDITIONS OR EVENTS BEFORE THE EVALUATION THAT CAN INFLUENCE SELECTION OF A DESIGN

Condition/Event	Implications/Alternatives
No comparison group available within the school/grade range	Use a time-series, longitudinal design (Design #1-2). Find a comparison group at another similar school. Use a test norm for comparison (Designs 5, 6). Create an appropriate criterion-referenced measure and compare performance against a set level of mastery (Design # 4).
No appropriate instruments available to measure achievement for program language group	Develop criterion-referenced instruments (which may be normed on a sample population) and determine an acceptable level of mastery. Use multiple measures of achievement: e.g., teacher judgment of proficiency on specific criteria.
No demographic data available on the student population	Develop survey instruments and administer. Use existing records (cumulative files, attendance records) to develop a profile. Select evaluation designs which do not require a comparison population (longitudinal designs, comparison with a local norm).
No baseline or pretest data available	Begin testing immediately and take a series of measures during the year (need an instrument not sensitive to testing effects). Base standards for success on multiple criteria of which test results are one. Use test norms as a means of comparison if norm group is similar. See designs 1-6.
Two or more language groups within the program	Evaluate each separately using most appropriate design(s).
No program goals or standards available	Identify implicit goals in law and policy. Look at needs assessment data and use community, professional and expert input to define standards. Use test data to specify standards, such as those specified by minimal competency program.
Tests are prescribed by an outside office or agency	Determine: the applicability of norms to the specific population. If not applicable, set standards in terms of other measures: a CRT, other tests, behavior.

Condition/Event	Implications/Alternatives
More than one level of achievement or language competence per grade	Set standards by language proficiency level rather than by grade.
Small number of students in the bilingual program	Compare outcomes with a prescribed mastery level on a CRT (Design 4). Use multiple standards. Use a longitudinal study or individual case-study approach (Design 1-2), especially where programs are individualized. Could individually match students with others at another school and compare progress (Designs 7, 12, 13).

EVENTS DURING THE PROGRAM THAT CAN AFFECT DESIGN IMPLEMENTATION

Event	Implications/Alternatives
Attrition of program students	If attrition is large, this may invalidate the planned comparison, since the compared groups are less similar. Instead, may have to use a prescribed mastery level for comparison (Design 4). If required to continue the evaluation, no generalizations of results should be made.
Influx of new students into bilingual program after it has been under way	Get demographic and other data (achievement data from past records if possible) to determine how similar the new students are to the current population. If different on identifiable variables, conduct a separate evaluation, for each group.
Classrooms or teachers reassigned causing a gain or loss in number of treatments within the bilingual program	Check comparison groups to maintain similarities. If comparing treatments, document the process and objectives of those remaining (Design 3). May have to drop or modify the comparisons planned. Reassignment during the year may cause temporary setback in learning: "pretest" in new assignments as soon as possible to determine status of new classes. May have to compare outcomes with a prescribed mastery level on a CRT and drop comparisons with dissimilar groups (Design 4).
During the year students are transitioned from native language to English	If they remain in the same bilingual classes, evaluation may be continued. If they are changed, causing attrition from the program, posttest as if in the program but perform their comparison separately.
Program expansion possible: More money and other resources available for program and evaluation	Add an evaluation of variations in treatment within the program (Design 3). Collect additional demographic data which may help in finding causes of unexpected outcomes. Be sure to document any changes in the program as a result of expansion; check comparison groups to insure they remain similar.

Events During the Program (continued)

Event	Implications/Alternatives
Program begins during the school year: no pretest data available.	Test as soon as possible. Can use data used as basis of selection into the program (e.g. score on English proficiency test) as one of multiple measures. If program begins during the year, you then have a unique measure of how children did in absence of the program: use achievement indicators taken earlier as a point of comparison, but document the differences between programs well.
Students were mislabeled, discovered later.	If students are retested, must consider intervening instruction in determining level used as baseline or pre-program. These students may not be appropriate to include in the planned comparison: conduct a separate evaluation of their performance.

EVENTS AT THE END OF THE PROGRAM WHICH MAY AFFECT USE OF PROGRAM DATA

Event	Implications/Alternatives
Incomplete data received on the posttest	Go back and retest to get missing data if possible. (Particularly important if have few students in the program.) Otherwise, analyze data but look for variables which make students without posttest data different from those with complete data.
Different instrument used for posttest	Check to insure comparability of pre and posttest instruments. If not comparable, planned analysis may be invalid. Tests should be same difficulty, normed on similar population, test same content. If instruments are comparable, need to insure that socres are equivalent.
No changes observed (or negative change observed) from pre- to posttest	Test may not be measuring areas of real change. Check instruments to insure they test content of bilingual curriculum. Use multiple standards (possibly including attitude, awareness of culture, attendance, etc.) to find more areas where bilingual program may have impact. Check how comparison populations differ to find alternative explanations. Look at differences in test performance of subgroups of students.
Scheduling problems: Posttesting conflicts with other required year-end testing	Try to overcome scheduling difficulties. Otherwise, test as close to the year end as possible (or during months specified by test manual).
Posttest population differs from pretest population	Collect demographic data to see where areas of difference occur. (Examine length of time in bilingual program; if too short may not get program effects.) If populations are similar, may cautiously infer that pretest scores were similar (if required to answer mandated evaluation question). Better to note changes; use a Mastery level design (or several if have definite subgroups). May try to trace students who left the program to infer their performance from more current records. Or, if have several measures taken during the year, can perform analyses for shorter periods during the year (e.g., October to February) when population was more stable.

Special Purpose Designs

This final section of the workshop presents designs that have been frequently used in research and evaluation studies and that may apply in bilingual settings if special conditions are met. They are presented in the same format as the recommended designs.

While at first glance many of the designs appear to be inappropriate for bilingual programs because random assignment is not permitted, matching impossible, and withholding the treatment illegal, some of these special purpose designs may be incorporated to answer sub-sets of evaluation questions that focus on research issues. For example, your program staff may wish to know which of two alternative reading programs best promotes program goals. Under such a direction, you may have permission from the local education agency to randomly assign materials to students or classes within the program. The variety of designs applicable to bilingual evaluations expands somewhat as evaluators are encouraged to answer basic questions about learning in bilingual settings.

Before referring to the models, be sure to review the conditions that need to be met for these special purpose designs. These conditions are summarized in Figure 2-7.

Figure 2-7

Special Purpose Designs for Bilingual Programs

Evaluation Purpose	Design	Implied Comparison	Requirements of the Design
Show program effectiveness by comparing to alternative or regular program	Non-equivalent control group	Same/similar students and program	(copy the requirements from attached)
Show program effectiveness in terms of "regular" program	Regression	To students from same population	(copy)
Investigate research-hypotheses relevant to bilingual learning and instruction	Control group	To same students	(copy)
	Cohort	To age-grade peers	(copy)

Special Purpose: To demonstrate program effectiveness by comparing your program to alternative/regular instructional programs

Requirements

Non-equivalent Control Group: (Design 7)

1. You must find a group of students who are similar in all characteristic to your group that are thought to contribute to program goals.

2. You must find these students in programs with similar goals to yours or you will be comparing apples and oranges.

CAUTION

Be wary of like-appearing groups because surface similarities, such as ethnicity or language dominance can hide - other important differences such as ses or level of proficiency in English and the native language. In practice, the likelihood that you will find two programs with similar goals, strategies, materials, and students is so low that you may wish to use this design for making within program comparisons such as between two versions of the same program.

Regression Projection and Discontinuity: (Designs 8-9)

1. Students must be placed in the bilingual program, and excluded from it on the basis of a composite index or pretest score.

2. Students in the bilingual program must be samples from the same population as students who are kept out of the program. They differ only in the respect that they scored lower on the placement index or pretest.

CAUTION

Most bilingual programs neither use a single pretest score nor a composite index for placing students; thus, you will not have the necessary information for these designs. A more serious drawback to these designs is the assumption that bilingual students form a sub-sample of the students taking the pretest who differ from the general sample only with respect to their lower test score. In most institutions, bilingual students are different along a variety of dimensions, ethnicity, acculturation, language proficiency and cannot be properly said to be drawn from the same population. This can occur when a program is located in a racially isolated school and not all students can be served or when program selection is based on a single variable, e.g., English test score. Since the "same population" assumption is almost routinely violated, this should be implemented with great caution.

Requirements

Control Group Designs: (Designs 10-11)

1. These designs require random assignment. This is possible when you have several bilingual classrooms at the same grade level with the same objectives and can then randomly assign students to classes. You might also randomly assign students to particular materials used in one classroom or to particular instructional modules that have different language instruction strategies.

2. This design is possible only when program objectives are comparable at the unit of interest, such as the classroom or individual student level.

CAUTION

These designs are possible if you can limit the scope of the research questions being asked to small groups or classrooms.

Cohort Designs: (Designs 12-13)

1. These designs are most useful for newly-installed bilingual programs where you are comparing like students at the same grade level but who are passing through the school system at different times.

2. Cohort designs are possible also if you have an articulated bilingual program with comparable goals and are making the comparisons using longitudinal data.

CAUTION

As mentioned above, these are most powerful when you are first installing a program and the "cohorts" are students who have not been in the bilingual program.

Intact Group Designs

Evaluation purpose:	Determine program effectiveness
Typical question:	How well are students in the program doing compared to what they would have done in an alternate program?
Comparison:	Non-equivalent group

Why This is Important: Requirements for comparative information are difficult to fill, but if similar student populations enrolled in similar programs with like objectives can be located, you can rule out many competing explanations for program effects.

Who Serves as the Sample?

Intact classrooms of students who represent the same population ethnically, linguistically, and socio-economically as your program students.

What Kinds of Data Should be Gathered?

In addition to student achievement data, a general requirement for comparisons between groups, you will need to document the similarity of program goals and the differences between the bilingual programs that make these comparisons between "alternative" treatments. You will need to look carefully at pretest scores on all measures to establish the a priori equivalence of the comparison groups.

What are the Threats to Validity Encountered in These Designs?

Sampling is the most obvious threat; your groups, despite your best efforts, may be very different.

Model 7: Non-equivalent Control Group Design

The non-equivalent control group design is diagrammed as follows:

	Time		
	1 (pre)		2 (post)
Experimental Group	O	X	O
Control Group	O		O

Description

1. The first step is to identify a comparison group that approximates the experimental groups as closely as possible.

1A. This would include grade level, socioeconomic status, pretest scores, language proficiency, age and sex, as well as other significant educational characteristics.

1B. The similarities between the two groups must be well documented to substantiate conclusions related to program effects.

2. Both groups should be pretested at approximately the same time, using the same instrument.

2A. The pretest scores of the two groups serve as one very important check on their comparability.

2B. Widely different pretest scores between the groups indicate that they are not equivalent on the variable with the highest correlation to program outcome measures. Any comparisons between the two groups with significantly different pretest scores should report this difference and and make statistical adjustments if possible.

3. During the course of the treatment, both groups should be monitored to insure that they are not"contaminated" by any unusual events or unexpected instructional changes, and that program X was indeed implemented.

4. At the end of the treatment, the test should be administered to both groups at the same time.

5. The difference between the two groups in their average growth from pretest to posttest serves as the measure of the effectiveness of the program.

6. Because the two groups often are not identical on the pretest, an analysis of covariance is frequently used to adjust for initial group differences so that relative growth of the two groups can be compared.

Advantages of Model # 7

Because random assignment is not required, this design is more practical in school settings.

Disadvantages of Model # 7

1. It is not possible to completely rule out the influence of selection factors in the final results.
2. Identifying and documenting the similarity of the experimental and comparison groups can be time consuming and complicated if relevant information has not already been collected.
3. Matching groups on all relevant educational variables is most difficult. There is always the possibility that some variable related to program success has not been identified.
4. In the case of bilingual programs, it may not even be possible to find a similar group for comparison.

Data Analysis References

Tatsuoka, M. Significance Tests: Univariate and Multivariate. Champaign, Illinois. Institute for Personality and Ability Testing, 1971.

Regression Designs

Evaluation Purpose: To show program effectiveness in terms of the "regular" program

Typical Question 3: How well are students in the program doing compared to English proficient students?

Comparison: Non-Equivalent Group

Why This is Important: Because of the difficulties in locating groups for comparison purposes, the best comparison available may be one that compares students in the program with those who scored too high to qualify for the program. Also, since the aim of many bilingual programs is to promote the students' ability to function in monolingual settings, comparison with students whose language proficiency is deemed "adequate" is a valid one.

Who serves as the sample?

Students in the bilingual program are compared to students who took the same placement test but who scored above the cut off point. The best comparison groups are the "borderline" students, those clustering just above and just below the cutoff points used to determine bilingual program enrollment. These students are likely to be similar in every way, except for a few points difference in test score. The comparison may be done in two ways:

1. By comparing the size of the difference in posttest performance between the two groups at the pretest cut off score or

2. By estimating the bilingual students' average posttest score from their pretest performance and comparing it with this actual posttest performance.

The two design models for comparing those qualifying for the bilingual program with those who "test out" are:

C. Regression discontinuity

9. Regression projection

What kinds of data should be gathered?

It is essential that test data from the selection instrument(s) be available, as well as posttest data. Any number of screening and selection instruments and strategies can be employed, so long as they can be combined into one composite number representing a distinct cut off score. If the same test is not used for "pre" and "post" measures, the posttest must be highly correlated with the measures used for placement.

In addition, a documentation of program processes using observation, student and teacher report, videotape and written plans and descriptions is helpful to determine if the bilingual program did, in fact, exist. Data on potential differences between the children that tested out of the program and those selected are helpful to explain differences between the two groups that may not be attributable to participation in the bilingual program.

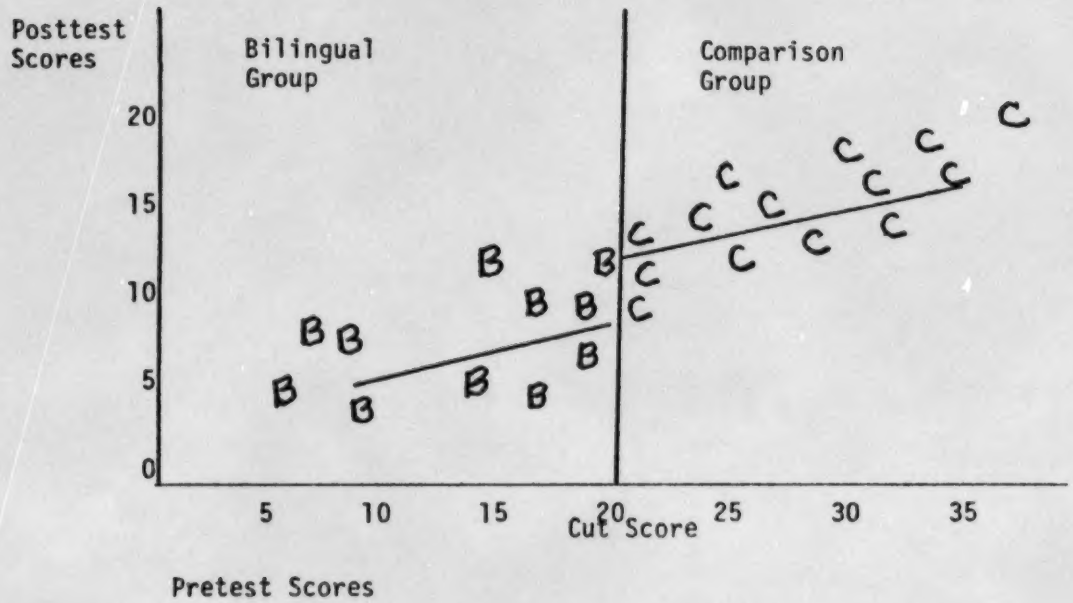
What are the threats to validity encountered in these designs?

The two most important threats to the validity of conclusions based on non-equivalent group comparisons are due to sample and instrumentation:

1. The two groups identified for the evaluation are different. The comparison group consists of students who score higher on the selection instrument(s), while the experimental group scores the lowest.

2. The comparison between groups assumes that students' growth and program between pretest and posttest are linear across all scores and parallel for the two groups. It may be that students scoring at the extreme low end show different growth patterns than those scoring near the cut-point.
3. If there are differential growth rates associated with second language acquisition, and these rates are non-linear, or differ according to student population, the regression models are inappropriate.
4. The comparison hinges on the validity of the selection procedure for separating the two groups. The comparisons are questionable if students are placed into the program who should not be or if students "tested out" of the program but did so by chance. Further, regression analyses are <u>inappropriate</u> if the students selected for the program are from a different population (e.g., foreign students all score low and Anglos high on the selection test) than students testing above the cut score for the program.

Model 8 Regression Discontinuity Design



Description

1. Students are given a pretest or are subject to a set of selection criteria for which a cut-point is established that sorts students into the bilingual program.

2. At the conclusion of the program, all students taking the pretest or involved in the selection take a posttest.

3. The comparison is made by using regression analysis on both group's pre and post scores, so that a regression line of best fit is plotted for each group.

4. The regression line representing each group's best pre-posttest score "fit" is extended for each group until it reaches the cut-point.

5. The effectiveness of the program is determined by the difference between the regression lines at the cut-off score point. Tallmadge and Horst (1974) suggest that if the group scoring below the cut off point has a regression line that is over 1/3 of a standard deviation above the no-treatment group's pretest score, then the difference between the two groups is educationally significant.

Advantages of Model #8:

1. Multiple selection criteria can be used to assign students to the program, as long as a composite cut-score can be established.
2. The model serves in evaluations where all eligible students must be enrolled in the program.

Disadvantages of Model #8:

1. It requires that a numerical cut score be set and that students below the score go into the bilingual program, and those above be kept out. It does not allow for students who score above the cut point to be enrolled in the program, even if other factors might suggest that be so.
2. When comparing LES/NES students to non-program participants, it is not possible to meaningfully define a population where half is assigned to the treatment and half stays out when the criteria for eligibility is qualitative. Thus, Models 6-7 have little application in these instances.

3. The comparison assumes that the skill being tested, in most cases language proficiency, grows in a linear fashion for both monolinguals and second language learners. There is no evidence that growth patterns in language skills are the same for LES/NES and FES students or that the program has the same (linear and additive) influence on all participants.
4. The comparison requires some mathematical sophistication and access to a computer with one of the standard regression analysis programs.

Data Analysis References:

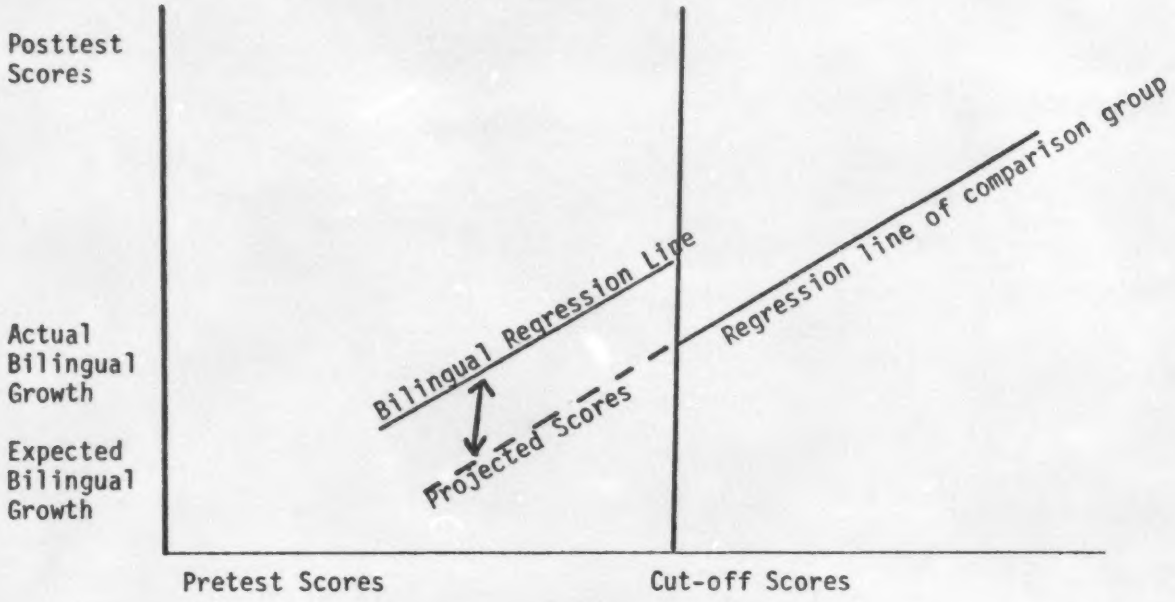
Cook, T., & Campbell, D. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally College Publishing Co., Chicago, 1979.

Horst, D.G., Tallmadge, G.K., & Wood, C. A Practical Guide to Measuring Project Impact on Student Achievement. RMC Research Corporation, Los Altos, CA, 1974.

Tallmadge, G. K., & Wood, C. User's Guide: ESEA Title I Evaluation and Reporting System. RMC Research Corporation, Mountain View, CA, 1974

Kerlinger, F. & Pedhazur, E. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, 1973.

Model 9: The Regression Projection Design



Description

1. Students are given a pretest or subject to a set of selection criteria for which a cut-point is established that sorts students into the bilingual program.

2. At the conclusion of the program, both program and non-program students take the same posttest.

3. Regression lines are constructed for both groups using the pre and posttest scores for both groups.

4. Instead of comparing the difference between how the bilingual and regular program regression lines intercept the cut-point line, the regression projection model projects the line of the comparison group across the cut-off line through the treatment group distribution.

5. This projected line represented the expected performance of the bilingual students, if they had not participated in the program.

In other words, the performance of the students who were not participating in the treatment is used to estimate what the performance of the program students would have been without the program. This "no treatment expectation" is compared with the performance of students in the bilingual program, as measured by their regression line to determine program effectiveness.

Advantages of Model 9

1. This design has the same advantages as Model 6.

Disadvantages of Model 9

1. This design has the same disadvantages as Model 6.

2. Like Model 6, it is more dependent on the assumption that student growth is even across all scores.

3. This model, like Model 6, is not likely to find program effects when students have different performance levels based on differential program effects for individual students (aptitude treatment interactions).

Data Analysis References

Cook, T., & Campbell, D. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally College Publishing Co., Chicago, Ill. 1979.

Horst, D.G., Tallmadge, G.K. & Wood, C. A Practical Guide to Measuring Project Impact on Student Achievement. RMC Research Corporation, Los Altos, California, 1974.

Tallmadge, G.K. & Wood, C. User's Guide: ESEA Title I Evaluation and Reporting System. RMC Research Corporation, Mountain View, California, 1976.

Equivalent Group Designs

Evaluation Purpose: To answer "research" questions within a bilingual program.
Question 1: How well are students in the program doing compared to what they would have done without the program or in a different program?
Comparison: Equivalent Group

Why Is This Important? The answer to this question is one frequently required by most regulatory agencies. They want to know if the program is justified; is it doing something for the students that they would not have received otherwise? Is the program helping the students, or is their achievement the same as it would have been if the bilingual program did not exist?

Such results must be included by law in a bilingual program evaluation in several states. California, Minnesota and Texas are three states where programs must be evaluated thoroughly for their effects. (See R. Irizarry, 1978, for a complete legislation review.)

It is important to note that this question can never be answered as posed. Students can only be in one place at one time -- we can never know what their performance would have been in absence of the program. So we look for a comparison group that is as similar as possible in order to estimate what participants' performance might have been. This is a problem in bilingual education, where by law all needy students must be given special assistance. Therefore, any comparison group will differ on several significant variables, including English and native language proficiency, and program received. That is why documentation of program process is essential, and evaluators should not imply that they can directly answer this evaluation question.

Who serves as the sample?

In order to show that students benefitted from the program, you need to select groups that either are not enrolled in the program or are in different programs and compare their performance to that of students in your program. If your students outperform those in the "control" or "comparison" group, then you may infer that they did better than they would have had there been no bilingual program. The key to collection of outcomes data for this question is the selection of an appropriate comparison group. Thus, designs for answering the "in the absence of treatment" question are organized around the kinds of groups that you

choose to compare with your program. Since you cannot infer that students benefitted from the program unless these students were compared with similar pupils, the designs are loosely called "equivalent group" designs. Equivalence in this instance means either that students were randomly assigned to program or no-program so that differences in population are evened out, or that they were matched on important variables.

The five equivalent-group comparisons presented in this section include:

Model 10: Pretest-posttest control group design

11: Posttest-only control group design

12: Single cohort comparisons

13: Multiple cohort comparisons

What kinds of data should be gathered?

Although you will collect data on student achievement in these designs, any behavior that can properly be labelled "program outcomes" is eligible. This would include measures of student self-concept, attitude to program, degree of parent participation, teacher attitude toward students' native language, or improved ethnic group relations within the school.

In addition, there are many other kinds of data that may be helpful in interpreting equivalent group comparisons. If you are comparing students in two different bilingual programs (or even classrooms), it is helpful to document the methods to be compared before the onset of the program (or program year). This requires that teachers sit down

beforehand and state plainly what they plan to do. Such plans should include: a) percentage of instruction in each language and by whom; b) division of the school day by subject, content and mode of instruction; and c) an estimate of how the curriculum will match the needs of assigned pupils. Such documentation will point out real differences in seemingly similar programs, or help identify why different programs have different outcomes.

Answering outcome questions about programs will inevitably also need to consider processes. While outcomes are being compared, it is always helpful to look at differences in treatments. Thus, use of this design should be accompanied with observation and interview assessment techniques, as well as testing.

What are the threats to validity encountered in these designs?

The following are the major validity threats which pertain to these designs:

a) Observed differences could be attributable to differences in the student population. Therefore you must either randomly assign students to classes, or examine background data to assure similarity.

b) Since the pupils in the two groups are essentially the same, it is important to avoid informing the treatment group that they are part of an experiment, thus creating Hawthorne effects.

c) The number of students within each treatment group must be large enough for statistical comparisons. Minimally, two or more classes could make up each treatment condition.

Model 10: The Pretest-Posttest Control Group Design

This design is presented in the following diagram:

	Time			
		1 (pre)		2 (post)
Experimental Group	R	O	X	O
Control Group	R	O		O

Description

1. Students for whom program X is appropriate are randomly assigned to form two groups, one which will receive program X and one which will not.

2. The group that does not get program X will participate in normal school activities, or if the evaluation is comparing two alternative programs, it will receive program Y.

3. Both groups are pretested at the same time using the same instrument, and sometime later (usually at the end of the school year) both are posttested.

3A. The average pretest score of the two groups should be the same.

3B. If it is significantly different, this suggests that the randomization process was imperfect, that some external event affected one group and not the other, or that some external event affected the test scores.

3C. If the pretest scores differ, comparisons between the two groups cannot be made until a statistical adjustment is done. A comparison between the average posttest scores will then reflect the relative benefits of program X.

Advantages of Model #10

1. Results are not subject to any of the competing explanations of selection, or extraneous factors discussed in the section on threats to validity.
2. Results are easy to interpret and explain.
3. Clear and distinct comparisons between educational alternatives sharing the same objectives are possible.
4. Using both pretests and posttests provides comparative measures between program X and the regular program, as well as some absolute measure of the growth of students participating in the alternative programs. (This will not be true in Model #2.)
5. Use of a pretest provides a check on the adequacy of randomization procedures and on pre-program group equivalence.

Disadvantages of Model #10

While this design has no conceptual flaws, it has some practical disadvantages.

1. In many circumstances (particularly in the case of bilingual programs), all students who are eligible for the program must be served. Thus, no control group can be used.
2. It is extremely rare that program alternatives exist so that students can be randomly assigned to "equivalent" treatments.

Data Analysis References

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Model 11: The Posttest-Only Control Group Design

This design is presented in the following diagram:

	Time		
	1 (pre)		2 (post)
Experimental Group	R	X	O
Control Group	R		O

Description

1. Students for whom program X is appropriate are randomly assigned to two groups, one which will receive program X and one which will not.

2. The group which does not get program X will participate in regular school activities or, in the case of comparison between two alternative programs, will receive the second program.

3. At the conclusion of the program both groups are given a posttest.

4. A comparison between the average scores of the two groups reflects the relative success of the programs.

Advantages of Model #11

1. This design is useful when a pretest might interfere with program effects in some manner, when a pretest is not available or when it would take too much time.
2. The random assignment initially equates the two groups so it is possible to wait until the end of the program before selecting all instruments.
3. Results are not subject to any of the threats to validity discussed in the preplanning section unless the two groups, by chance, are truly different in make-up.

Disadvantages of Model #11

While there are no conceptual drawbacks to this model, it has the same practical disadvantages as Model #1.

1. In many circumstances (including bilingual programs), all eligible students must be served, so no control group can be formed.
2. It is extremely rare that students can be randomly assigned to program alternatives.
3. By eliminating the pretest, this design does not yield any measure of student gain during the course of the program.
4. Without a pretest, there is no evidence that randomization produced initially equivalent groups. Poor randomization may threaten the validity of your conclusions.

Summary

Although the strongest designs allow the strongest inferences, it is rarely possible to structure the school situation to fit the models. However, Models 10-11 are not completely unattainable ideals. Under circumstances where an evaluator is asked to compare two alternative programs, he/she may be able to make use of these designs. Also, they can sometimes form a powerful part of a multiple design strategy in more extensive evaluations where students can be randomly assigned to sub-programs and where intra-program comparisons can be made.

Data Analysis References

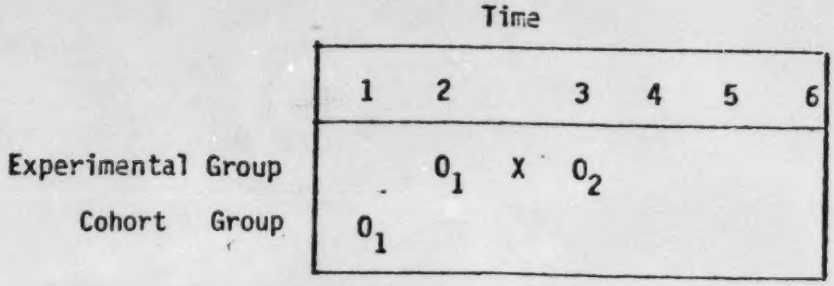
Whitla, D. K. (ed). Handbook of Measurement and Assessment in Behavioral Sciences. Reading, Massachusetts: Addison-Wesley, 1968.

Kerlinger, F. (2nd ed), Foundations of Behavioral Research. New York: Holt, Rinehart and Winston, 1973.

Lohnes, P., & Cooley, W. Introduction to Statistical Procedures with Computer Exercises. New York: Wiley, 1968.

Model 12: Single Cohort Comparisons

A cohort comparison design would be diagrammed as follows:



Description

1. Cohort comparisons use students from the previous year as the comparison group. The evaluator selects students from the same grade as those who are to be in the bilingual program, but who were enrolled in school a year before the program began.

2. Students in the selected grade are tested at the end of the year. This test serves as an estimate of what students in the bilingual program would have done without the program.

Model 13: Multiple Cohort Comparisons

A multiple cohort comparison using one group from the past and one group from the future is diagrammed below:

	Time					
	1	2	3	4	5	6
Cohort Group	Grade 5 O	O				
Experimental Group			Grade 6 O	X	O	
Cohort Group					Grade 7 O	

Description

1. Students who are both one grade behind and one ahead of bilingual students are selected for comparison.

2. These students must not have been in the program and should be similar to program students.

Advantages of Models #12-13

1. They can be used in circumstances where contemporaneous comparison groups cannot be found.
2. The results of the analysis are easy to understand and to present.
3. If the school population is relatively stable, then the group of students who passed through a certain grade a year ago will resemble the group of students who are in that grade at the present time. They may even have the same teachers and classroom aides. As a result, cohort groups can be among the most equivalent of all the non-randomly assigned comparison groups.

Disadvantages of Models #12-13

1. Care must be taken in both simple and multiple cohort designs to insure that the measures have test-retest reliability for 6 month periods or longer.

2. Cohort comparisons are limited to circumstances where programs are just being initiated.

3. The evaluator must be particularly careful to establish the similarity of the school population during the time of the study. Documenting this fact is an important part of the evaluation, and helps to strengthen the validity of the results.

4. In communities where an influx of Non-English Speaking families have made bilingual programs a necessity, cohort designs may be completely inappropriate. Stable multilingual communities must be differentiated from changing multilingual communities when deciding to implement this type of design.

Data Analysis References

Cook, T., & Campbell, D. Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand McNally Publishing Company, Chicago, 1979.

References

Andrews, F., Klem, L., Davidson, T. et al. A Guide for Selecting Statistical Techniques for Analyzing Social Science Data. Institute of Social Research, University of Michigan, Ann Arbor, Michigan, 1978.

Alkin, M., Daillak, R., & White, P. Using Evaluations. Sage Publications, Beverly Hills, 1979.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand-McNally, 1963, 171-246.

Cappell, F., Sisson, T., & Winters, L. Final report: Basic skills Learning Centers Evaluation. Center for the Study of Evaluation, September, 1979.

Cronbach, L. T. Designing educational evaluations. Occasional Papers of the Stanford Evaluation Consortium. April, 1978.

Cook T., & Campbell, D. Quasi-experimentation: Design and analysis issues for field settings. Rand-McNally College Publishing Company, Chicago, 1979.

Horst, D. G. Guide to measuring project impact on student achievement. RMC Research Corporation, Los Altos, California, 1975. Government Printing Office Stock Number 017-080-01460.

Kerlinger, F., & Pedhazur, E. Multiple Regression in behavioral research. Holt, Rinehart and Winston, New York, 1973.

Linn, R. L., & Slinde, J. A. The determination of significance of change between pre and posttesting periods. Review of Educational Research, 1977, 47(1), 121-150.

Morris, L. Program Evaluation Kit. Sage Publications, Beverly Hills, 1978

Patton, M. Utilization-Focused Evaluation. Sage Publications, Beverly Hills, 19 .

Popham, W. J. Criterion Referenced Measurement. Englewood Cliffs, New Jersey. Prentice Hall, 1968.

Tallmadge, G. K., & Wood, C. User's guide: ESEA Title I evaluation and Reporting System. RMC Research Corporation, Mountain View, California, 1976.

Tallmadge, G. K., & Horst, D. A procedural guide for validating achievement gains in educational projects. RMC Research Corporation, Mountain View, California, 1976. Government Printing Office Stock Number 017-080-01516-1.

List of Figures

2-1 Events in the Life of a Bilingual Program Affecting Design

2-2 Steps in selecting evaluation designs

2-3 Implied comparisons

2-4 Requirements of typical design models: A reference

2-5 Relative advantages of typical evaluation designs

2-6 Recommended designs for bilingual programs

2-7 Special Purpose Designs